Equi-energy sampler with applications in statistical inference and statistical mechanics

The Annals of Statistics
2006, Vol. 34, No. 4, 1581–1619
DOI: 10.1214/009053606000000515

DISCUSSION PAPER

EQUI-ENERGY SAMPLER WITH APPLICATIONS IN STATISTICAL INFERENCE AND STATISTICAL MECHANICS[1,2,3]

BY S. C. KOU, QING ZHOU AND WING HUNG WONG

Harvard University, Harvard University and Stanford University

We introduce a new sampling algorithm, the equi-energy sampler, for efficient statistical sampling and estimation. Complementary to the widely used temperature-domain methods, the equi-energy sampler, utilizing the temperature–energy duality, targets the energy directly. The focus on the energy function not only facilitates efficient sampling, but also provides a powerful means for statistical estimation, for example, the calculation of the density of states and microcanonical averages in statistical mechanics. The equi-energy sampler is applied to a variety of problems, including exponential regression in statistics, motif sampling in computational biology and protein folding in biophysics.

1. Introduction. Since the arrival of modern computers during World War II, the Monte Carlo method has greatly expanded the scientific horizon to study complicated systems ranging from the early development in computational physics to modern biology. At the heart of the Monte Carlo method lies the difficult problem of sampling and estimation: Given a target distribution, usually multidimensional and multimodal, how do we draw samples from it and estimate the statistical quantities of interest? In this article, we attempt to introduce a new sampling algorithm, the equi-energy sampler, to address the problem. Since the Monte Carlo method began from calculations in statistical physics and mechanics, to introduce the equi-energy sampler, we begin from statistical mechanics.

The starting point of a statistical mechanical computation is the energy function or Hamiltonian $h(x)$. According to Boltzmann and Gibbs, the distribution of a system in thermal equilibrium at temperature T is described by the Boltzmann distribution,

$$p_T(x) = \frac{1}{Z(T)} \exp\bigl(-h(x)/T\bigr), \tag{1}$$

Received June 2004; revised March 2005.

[1]Supported in part by NSF, NIH and Harvard University Clarke–Cooke Fund.

[2]Discussed in 10.1214/009053606000000470, 10.1214/009053606000000489, 10.1214/009053606000000498 and 10.1214/009053606000000506; rejoinder at 10.1214/009053606000000524.

[3]S. C. Kou and Qing Zhou contributed equally to this work.

AMS 2000 subject classifications. Primary 65C05; secondary 65C40, 82B80, 62F15.

Key words and phrases. Sampling, estimation, temperature, energy, density of states, microcanonical distribution, motif sampling, protein folding.

where $Z(T) = \sum_x \exp(-h(x)/T)$ is referred to as the partition function. For any state function $g(x)$, its expectation $\mu_g(T)$ with respect to the Boltzmann distribution is called its Boltzmann average, also known as the thermal average in the physics literature,

$$\mu_g(T) = \sum_x g(x) \exp\bigl(-h(x)/T\bigr)/Z(T). \tag{2}$$

To study the system, in many cases we are interested in using Monte Carlo simulation to obtain estimates of Boltzmann averages as functions of temperature for various state functions. In addition to Boltzmann averages, estimating the partition function $Z(T)$, which represents the dependency of the normalization factor in (1) as a function of temperature, is also of significant interest, as it is well known that many important thermodynamic quantities, such as free energy, specific heat, internal energy and so on, can be computed directly from the partition function (see Section 4). The fundamental algorithm for computing Boltzmann averages is due to Metropolis et al. [29], who proposed the use of a reversible Markov chain constructed in such a way so as to guarantee that its stationary distribution is the Boltzmann distribution (1). Later, this algorithm was generalized by Hastings [11] to allow the use of an asymmetric transition kernel. Given a current state x, this Metropolis–Hastings algorithm generates a new state by either reusing the current state x or moving to a new state y drawn from a proposal kernel $K(x \to y)$. The proposal state y is accepted with probability $\min(1, MR)$ where MR is the Metropolis–Hastings ratio $p_T(y)K(y \to x)/p_T(x)K(x \to y)$. The algorithm in this way generates a Markov chain $X_i, i = 1, \ldots, n$. Under ergodic conditions [39], the time average $n^{-1} \sum_{i=1}^n g(X_i)$ provides a consistent estimate of the Boltzmann average (2).

The Metropolis algorithm, however, can perform poorly if the energy function has many local minima separated by high barriers that cannot be crossed by the proposal moves. In this situation the chain will be trapped in local energy wells and will fail to sample the Boltzmann distribution correctly. To overcome this problem one can design specific moves that have a higher chance to cut across the energy barrier (e.g., the conditional sampling moves in Gibbs sampling) or to add auxiliary variables so that the energy wells become connected by the added dimension (e.g., the group Ising updating of Swendsen and Wang [37], or the data-augmentation technique of Tanner and Wong [38]). However, these remedies are problem-specific and may or may not work for any given problem. A breakthrough occurred with the development of the parallel tempering algorithm by Geyer [8] (also called exchanged Monte Carlo; Hukushima and Nemoto [13]). The idea is to perform parallel Metropolis sampling at different temperatures. Occasionally one proposes to exchange the states of two neighboring chains (i.e., chains with adjacent temperature levels). The acceptance probability for the exchange is designed to ensure that the joint states of all the parallel chains evolve according to the Metropolis–Hastings rule with the product distribution (i.e., the product of the Boltzmann distributions at the different temperatures) as the target distribution. Geyer's initial

objective was to use the (hopefully) faster mixing of the high-temperature chains to drive the mixing of the whole system, thereby to achieve faster mixing at the low-temperature chain as well. It is clear that parallel tempering can provide estimates of Boltzmann averages at the temperatures used in the simulation. Marinari and Parisi [27] developed simulated tempering that uses just a single chain but augments the state by a temperature variable that is dynamically moved up or down the temperature ladder. These authors further developed the theory of using the samples in the multiple temperatures to construct estimates of the partition function, and to investigate phase transition. In the meantime, Geyer [9] also proposed a maximum likelihood approach to estimate the ratios of normalization constants and hence obtain information on the partition function at the selected temperatures.

In contrast to the statistical mechanical computations, the starting point in statistical inference is usually *one* given distribution, for example, the distribution on a high-dimensional parameter space. If we take the energy to be the negative log-density function in this case, we are then interested in obtaining the Boltzmann average only at $T = 1$. The Metropolis and related algorithms have been developed and applied to solve many statistical computation problems, and have greatly enhanced our ability to analyze problems ranging from image analysis to missing value problems to biological sequence analysis to single-molecule chemistry [6, 7, 17, 20, 22, 38]. However, as Geyer, Marinari, Parisi and others have pointed out, even if the immediate interest is at $T = 1$, simulation at temperatures other than $T = 1$ is often necessary in order to achieve efficient sampling. Furthermore, computing the normalization constants (i.e., partition function) is also important in statistical tasks such as the determination of likelihood ratios and Bayes factors [9, 10, 28].

It is thus seen that historically dynamic Monte Carlo methods were developed to simulate from the Boltzmann distribution at fixed temperatures. These methods aim to provide direct estimates of parameters such as Boltzmann averages and partition functions, which are functions of temperature. We hence refer to these methods as *temperature-domain* methods. The purpose of this article is to develop an *alternative* sampling and estimation approach based on *energy-domain* considerations. We will construct algorithms for the direct estimation of parameters such as microcanonical averages and density of states (see Section 2) that are functions of *energy*. We will see in Section 2 that there is a duality between temperature-domain functions and energy-domain functions, so that once we have obtained estimates of the density of states and microcanonical averages (both are energy-domain functions), we can easily transfer to the temperature domain to obtain the partition function and the Boltzmann averages. In Section 3 we introduce the equi-energy sampler (EE sampler), which, targeting energy directly, is a new Monte Carlo algorithm for the efficient sampling from multiple energy intervals. In Sections 4 and 5 we explain how to use these samples to obtain estimates of density of states and microcanonical averages, and how to extend the energy-domain method

to estimate statistical quantities in general. In Section 6 we illustrate the wide applicability of this method by applying the equi-energy sampler and the estimation methods to a variety of problems, including an exponential regression problem, the analysis of regulatory DNA motifs and the study of a simplified model for protein folding. Section 7 concludes the article with discussion and further remarks.

2. Energy–temperature duality. The Boltzmann law (1) implies that the conditional distribution of the system given its energy $h(x) = u$ is the uniform distribution on the equi-energy surface $\{x : h(x) = u\}$. In statistical mechanics, this conditional distribution is referred to as the microcanonical distribution given energy u. Accordingly, the conditional expectation of a state function $g(x)$ given an energy level u is called its microcanonical average:

$$\nu_g(u) = E\big(g(X)|h(X) = u\big). \tag{3}$$

Note that (3) is independent of the temperature T used in the Boltzmann distribution for X. Suppose that the infinitesimal volume of the energy slice $\{x : h(x) \in (u, u + du)\}$ is approximately equal to $\Omega(u)\,du$. This function $\Omega(u)$ is then called the *density of states* function. If the state space is discrete, then we replace the volume by counts, in which case $\Omega(u)$ is simply the number of states with the energy equal to u. Without loss of generality, we assume that the minimum energy of the system $u_{\min} = 0$. The following result follows easily from these definitions.

LEMMA 1. *Let $\beta = 1/T$ denote the inverse temperature so that the Boltzmann averages and partition function are indexed by β as well as by T; then*

$$\mu_g(\beta^{-1})Z(\beta^{-1}) = \int_0^\infty \nu_g(u)\Omega(u)e^{-\beta u}\,du.$$

In particular, the partition function $Z(\beta^{-1})$ and the density of states $\Omega(u)$ form a Laplace transform pair.

This lemma suggests that the Boltzmann averages and the partition function can be obtained through Monte Carlo algorithms designed to compute the density of states and microcanonical averages. We hence refer to such algorithms as energy-domain algorithms.

The earliest energy-domain Monte Carlo algorithm is the multicanonical algorithm due to Berg and Neuhaus [2], which aims to sample from a distribution flat in the energy domain through an iterative estimation updating scheme. Later, the idea of iteratively updating the target distribution was generalized to histogram methods (see [18, 40] for a review). The main purpose of these algorithms is to obtain the density of states and related functions such as the specific heat. They do not directly address the estimation of Boltzmann averages.

In this article we present a different method that combines the use of multiple energy ranges, multiple temperatures and step sizes, to produce an efficient sampling scheme capable of providing direct estimates of all microcanonical averages as well as the density of states. We do not use iterative estimation of density of states as in the multicanonical approach; instead, the key of our algorithm is a new type of move called the equi-energy jump that aims to move directly between states with similar energy (see the next section). The relationship between the multicanonical algorithm and the equi-energy sampler will be discussed further in Section 7.

3. The equi-energy sampler.

3.1. *The algorithm.* In Monte Carlo statistical inference one crucial task is to obtain samples from a given distribution, often known up to a normalizing constant. Let $\pi(x)$ denote the target distribution and let $h(x)$ be the associated energy function. Then $\pi(x) \propto \exp(-h(x))$. For simple problems, the famous Metropolis–Hastings (MH) algorithm, which employs a local Markov chain move, could work. However, if $\pi(x)$ is multimodal and the modes are far away from each other, which is often the case for practical multidimensional distributions, algorithms relying on local moves such as the MH algorithm or the Gibbs sampler can be easily trapped in a local mode indefinitely, resulting in inefficient and even unreliable samples.

The EE sampler aims to overcome this difficulty by working on the energy function directly. First, a sequence of energy levels is introduced:

$$H_0 < H_1 < H_2 < \cdots < H_K < H_{K+1} = \infty, \tag{4}$$

such that H_0 is below the minimum energy, $H_0 \le \inf_x h(x)$. Associated with the energy levels is a sequence of temperatures

$$1 = T_0 < T_1 < \cdots < T_K.$$

The EE sampler considers $K+1$ distributions, each indexed by a temperature and an energy truncation. The energy function of the ith distribution π_i $(0 \le i \le K)$ is $h_i(x) = \frac{1}{T_i}(h(x) \vee H_i)$, that is, $\pi_i(x) \propto \exp(-h_i(x))$. For each i, a sampling chain targeting π_i is constructed. Clearly π_0 is the initial distribution of interest. The EE sampler employs the other K chains to overcome local trapping, because for large i the energy truncation and the high temperature on $h_i(x)$ flatten the distribution $\pi_i(x)$, making it easier to move between local modes. The quick mixing of chains with large i is utilized by the EE sampler, through a step termed the *equi-energy jump*, to help sampling from π_i with small i, where the landscape is more rugged.

The equi-energy jump, illustrated in Figure 1, aims to directly move between states with similar energy levels. Intuitively, if it can be implemented properly in a sampling algorithm, it will effectively eliminate the problem of local trap. The

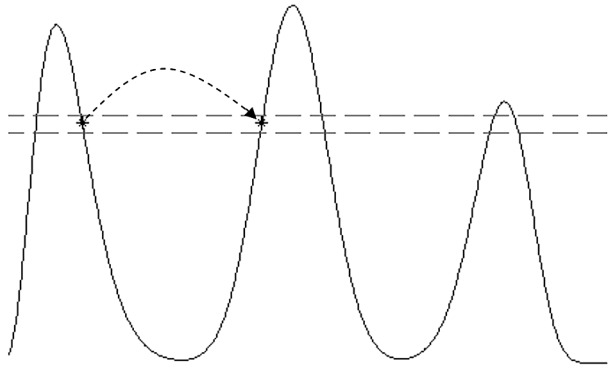

FIG. 1. *Illustration of the equi-energy jump, where the sampler can jump freely between the states with similar energy levels.*

fact that the ith energy function $h_i(x)$ is monotone in $h(x)$ implies that, above the truncation values, the equi-energy sets $S_i(H) = \{x : h_i(x) = H\}$ are mutually consistent across i. Thus once we have constructed an empirical version of the equi-energy sets at high-order π_i (i.e., π_i with large i), these high-order empirical sets will remain valid at low-order π_i. Therefore, after the empirical equi-energy sets are first constructed at high order π_i, the local trapping at low-order π_i can be largely evaded by performing an equi-energy jump that allows the current state to jump to another state drawn from the already constructed high-order empirical equi-energy set that has energy level close to the current state. This is the basic idea behind the EE sampler. We will refer to the empirical equi-energy sets as energy rings hereafter.

To construct the energy rings, the state space $\mathcal{X}$ is partitioned according to the energy levels, $\mathcal{X} = \bigcup_{j=0}^{K} D_j$, where $D_j = \{x : h(x) \in [H_j, H_{j+1})\}$, $0 \le j \le K$, are the energy sets, determined by the energy sequence (4). For any $x \in \mathcal{X}$, let $I(x)$ denote the partition index such that $I(x) = j$, if $x \in D_j$, that is, if $h(x) \in [H_j, H_{j+1})$.

The EE sampler begins from an MH chain $X^{(K)}$ targeting the highest-order distribution π_K. After an initial burn-in period, the EE sampler starts constructing the Kth-order energy rings $\hat{D}_j^{(K)}$ by grouping the samples according to their energy levels; that is, $\hat{D}_j^{(K)}$ consists of all the samples $X_n^{(K)}$ such that $I(X_n^{(K)}) = j$. After the chain $X^{(K)}$ has been running for N steps, the EE sampler starts the second highest-order chain $X^{(K-1)}$ targeting π_{K-1}, while it keeps on running $X^{(K)}$ and updating $\hat{D}_j^{(K)}$ $(0 \le j \le K)$. The chain $X^{(K-1)}$ is updated through two operations: the local move and the equi-energy jump. At each update a coin is flipped; with probability $1 - p_{\text{ee}}$ the current state $X_n^{(K-1)}$ undergoes an MH local move to give the next state $X_{n+1}^{(K-1)}$, and with probability p_{ee}, $X_n^{(K-1)}$ goes through an equi-energy jump. In the equi-energy jump, a state y is chosen uniformly from the

highest-order energy ring $\hat{D}_j^{(K)}$ indexed by $j = I(X_n^{(K-1)})$ that corresponds to the energy level of $X_n^{(K-1)}$ [note that y and $X_n^{(K-1)}$ have similar energy level, since $I(y) = I(X_n^{(K-1)})$ by construction]; the chosen y is accepted to be the next state $X_{n+1}^{(K-1)}$ with probability $\min(1, \frac{\pi_{K-1}(y)\pi_K(X_n^{(K-1)})}{\pi_{K-1}(X_n^{(K-1)})\pi_K(y)})$; if y is not accepted, $X_{n+1}^{(K-1)}$ keeps the old value $X_n^{(K-1)}$. After a burn-in period on $X^{(K-1)}$, the EE sampler starts the construction of the second highest-order [i.e., $(K-1)$st order] energy rings $\hat{D}_j^{(K-1)}$ in much the same way as the construction of $\hat{D}_j^{(K)}$, that is, collecting the samples according to their energy levels. Once the chain $X^{(K-1)}$ has been running for N steps, the EE sampler starts $X^{(K-2)}$ targeting π_{K-2} while it keeps on running $X^{(K-1)}$ and $X^{(K)}$. Like $X^{(K-1)}$, the chain $X^{(K-2)}$ is updated by the local MH move and the equi-energy jump with probabilities $1 - p_{\text{ee}}$ and p_{ee}, respectively. In the equi-energy jump, a state y uniformly chosen from $\hat{D}^{(K-1)}_{I(X_n^{(K-2)})}$, where $X_n^{(K-2)}$ is the current state, is accepted to be the next state $X_{n+1}^{(K-2)}$ with probability $\min(1, \frac{\pi_{K-2}(y)\pi_{K-1}(X_n^{(K-2)})}{\pi_{K-2}(X_n^{(K-2)})\pi_{K-1}(y)})$. The EE sampler thus successively moves down the energy and temperature ladder until the distribution π_0 is reached. Each chain $X^{(i)}, 0 \le i < K$, is updated by the equi-energy jump and the local MH move; the equi-energy move proposes a state y uniformly chosen from the energy ring $\hat{D}^{(i+1)}_{I(X_n^{(i)})}$ and accepts the proposal with probability $\min(1, \frac{\pi_i(y)\pi_{i+1}(X_n^{(i)})}{\pi_i(X_n^{(i)})\pi_{i+1}(y)})$. At each chain $X^{(i)}$, the energy rings $\hat{D}_j^{(i)}$ are constructed after a burn-in period, and will be used for chain $X^{(i-1)}$ in the equi-energy jump. The basic sampling scheme can be summarized as follows.

The algorithm of the EE sampler

Assign $X_0^{(i)}$ an initial value and set $\hat{D}_j^{(i)} = \varnothing$ for all i and j
For $n = 1, 2, \ldots$
  For $i = K$ downto 0
    if $n > (K - i)(B + N)$ do
    (* B is the burn-in period, and N is the
        period to construct initial energy rings *)
      if $i = K$ or if $\hat{D}^{(i+1)}_{I(X_{n-1}^{(i)})} = \varnothing$ do
        perform an MH step on $X_{n-1}^{(i)}$ with
            target distribution π_i to obtain $X_n^{(i)}$
      else
      with probability $1 - p_{\text{ee}}$, perform an MH step
          on $X_{n-1}^{(i)}$ targeting π_i to obtain $X_n^{(i)}$
      with probability p_{ee}, uniformly pick a state y from $\hat{D}^{(i+1)}_{I(X_{n-1}^{(i)})}$ and let

$X_n^{(i)} \leftarrow y$ with prob $\min(1, \frac{\pi_i(y)\pi_{i+1}(X_{n-1}^{(i)})}{\pi_i(X_{n-1}^{(i)})\pi_{i+1}(y)})$;
$X_n^{(i)} \leftarrow X_{n-1}^{(i)}$ with the remaining prob.
endif
if $n > (K-i)(B+N)+B$ do
$\hat{D}^{(i)}_{I(X_n^{(i)})} \leftarrow \hat{D}^{(i)}_{I(X_n^{(i)})} + \{X_n^{(i)}\}$
(* add the sample to the energy rings after the burn-in period *)
endif
endif
endfor
endfor

The idea of moving along the equi-energy surface bears a resemblance to the auxiliary variable approach, where to sample from a target distribution $\pi(x)$, one can iteratively first sample an auxiliary variable $U \sim \text{Unif}[0, \pi(X)]$, and then sample $X \sim \text{Unif}\{x : \pi(x) \geq U\}$. This approach has been used by Edwards and Sokal [5] to explain the Swendsen–Wang [37] clustering algorithm for Ising model simulation, and by Besag and Green [3] and Higdon [12] on spatial Bayesian computation and image analysis. Roberts and Rosenthal [32], Mira, Moller and Roberts [30] and Neal [31] provide further discussion under the name of slice sampling. Under our setting, this auxiliary variable (slice sampler) approach amounts to sampling from the lower-energy sets. In comparison, the EE sampler's focus on the equi-energy set is motivated directly from the concept of microcanonical distribution in statistical mechanics. More importantly, the equi-energy jump step in the EE sampler offers a *practical* means to carry out the idea of moving along the energy sets (or moving horizontally along the density contours), and thus provides an effective way to study not only systems in statistical mechanics, but also general statistical inference problems (see Sections 4, 5 and 6).

3.2. *The steady-state distribution.* With help from the equi-energy jump to address the local trapping, the EE sampler aims to efficiently draw samples from the given distribution π (which is identical to π_0). A natural question is then: In the long run, will the EE samples follow the correct distribution?

The following theorem shows that the steady-state distribution of chain $X^{(i)}$ is indeed π_i; in particular, the steady-state distribution of $X^{(0)}$ is $\pi_0 = \pi$.

THEOREM 2. *Suppose* (i) *the highest-order chain $X^{(K)}$ is irreducible and aperiodic*, (ii) *for $i = 0, 1, \ldots, K-1$, the MH transition kernel $T_{\text{MH}}^{(i)}$ of $X^{(i)}$ connects adjacent energy sets in the sense that for any j there exist sets $A_1 \subset D_j$, $A_2 \subset D_j$, $B_1 \subset D_{j-1}$ and $B_2 \subset D_{j+1}$ with positive measure such that the transition probabilities*

$$T_{\text{MH}}^{(i)}(A_1, B_1) > 0, \qquad T_{\text{MH}}^{(i)}(A_2, B_2) > 0$$

and (iii) *the energy set probabilities* $p_j^{(i)} = P_{\pi_i}(X \in D_j) > 0$ *for all* i *and* j. *Then* $X^{(i)}$ *is ergodic with* π_i *as its steady-state distribution*.

PROOF. We use backward induction to prove the theorem.

For $i = K$, $X^{(K)}$ simply follows the standard MH scheme. The desired conclusion thus follows from the fact that $X^{(K)}$ is aperiodic and irreducible.

Now assume the conclusion holds for the $(i+1)$st order chain, that is, assume $X^{(i+1)}$ is ergodic with steady-state distribution π_{i+1}. We want to show that the conclusion also holds for $X^{(i)}$. According to the construction of the EE sampler, if at the nth step $X_n^{(i)} = x$, then at the next step with probability $1 - p_{\mathrm{ee}}$ $X_{n+1}^{(i)}$ will be drawn from the transition kernel $T_{\mathrm{MH}}^{(i)}(x, \cdot)$, and with probability p_{ee} $X_{n+1}^{(i)}$ will be equal to a y from $\hat{D}_{I(x)}^{(i+1)}$ with probability

$$P\big(X_{n+1}^{(i)} = y\big) = \frac{1}{|\hat{D}_{I(x)}^{(i+1)}|} \min\bigg(1, \frac{\pi_i(y)\pi_{i+1}(x)}{\pi_i(x)\pi_{i+1}(y)}\bigg), \qquad y \in \hat{D}_{I(x)}^{(i+1)}.$$

Therefore for any measurable set A, the conditional probability

$$\begin{aligned}
&P\big(X_{n+1}^{(i)} \in A | X_n^{(i)} = x, X^{(i+1)}\big) \\
&\quad = (1 - p_{\mathrm{ee}}) T_{\mathrm{MH}}^{(i)}(x, A) \\
&\qquad + p_{\mathrm{ee}} \frac{1}{|\hat{D}_{I(x)}^{(i+1)}|} \sum_{y \in \hat{D}_{I(x)}^{(i+1)}} I(y \in A) \min\bigg(1, \frac{\pi_i(y)\pi_{i+1}(x)}{\pi_i(x)\pi_{i+1}(y)}\bigg) \\
&\qquad + p_{\mathrm{ee}} \bigg[1 - \frac{1}{|\hat{D}_{I(x)}^{(i+1)}|} \sum_{y \in \hat{D}_{I(x)}^{(i+1)}} \min\bigg(1, \frac{\pi_i(y)\pi_{i+1}(x)}{\pi_i(x)\pi_{i+1}(y)}\bigg)\bigg] I(x \in A).
\end{aligned}$$

Using the induction assumption of the ergodicity of $X^{(i+1)}$ and also the fact that the lower-order chain $X^{(i)}$ does not affect the higher-order chain $X^{(i+1)}$, we have, as $n \to \infty$,

$$\begin{aligned}
&P\big(X_{n+1}^{(i)} \in A | X_n^{(i)} = x\big) \\
&\quad = \int P\big(X_{n+1}^{(i)} \in A | X_n^{(i)} = x, X^{(i+1)}\big)\, dP\big(X^{(i+1)} | X_n^{(i)} = x\big) \\
&\quad = \int P\big(X_{n+1}^{(i)} \in A | X_n^{(i)} = x, X^{(i+1)}\big)\, dP\big(X^{(i+1)}\big) \\
&\quad \to (1 - p_{\mathrm{ee}}) T_{\mathrm{MH}}^{(i)}(x, A) \\
&\qquad + p_{\mathrm{ee}} \frac{1}{p_{I(x)}^{(i+1)}} \int_{y \in A \cap D_{I(x)}} \pi_{i+1}(y) \min\bigg(1, \frac{\pi_i(y)\pi_{i+1}(x)}{\pi_i(x)\pi_{i+1}(y)}\bigg) dy \\
&\qquad + p_{\mathrm{ee}} \bigg[1 - \frac{1}{p_{I(x)}^{(i+1)}} \int_{y \in D_{I(x)}} \pi_{i+1}(y) \min\bigg(1, \frac{\pi_i(y)\pi_{i+1}(x)}{\pi_i(x)\pi_{i+1}(y)}\bigg) dy\bigg] I(x \in A).
\end{aligned} \tag{5}$$

Similarly, as $n \to \infty$, the difference

$$P\big(X_{n+1}^{(i)} \in A | X_n^{(i)} = x, X_{n-1}^{(i)}, \ldots, X_1^{(i)}\big) - P\big(X_{n+1}^{(i)} \in A | X_n^{(i)} = x\big) \to 0.$$

Now let us define a new transition kernel $S^{(i)}(x, \cdot)$, which undergoes the transition $T_{\mathrm{MH}}^{(i)}(x, \cdot)$ with probability $1 - p_{\mathrm{ee}}$, and with probability p_{ee} undergoes an MH transition with the proposal density $q(x, y) = \frac{1}{p_{I(x)}^{(i+1)}} \pi_{i+1}(y) I(y \in D_{I(x)})$, that is, $\pi_{i+1}(y)$ confined to the energy set $D_{I(x)}$. We then note that the right-hand side of (5) corresponds exactly to the transition kernel $S^{(i)}(x, \cdot)$. Therefore, under the induction assumption, $X^{(i)}$ is asymptotically equivalent to a Markovian sequence governed by $S^{(i)}(x, \cdot)$.

Since the kernel $T_{\mathrm{MH}}^{(i)}(x, \cdot)$ connects adjacent energy sets and the proposal $q(x, y)$ connects points in the same equi-energy set, it follows from Chapman–Kolmogorov and $0 < p_{\mathrm{ee}} < 1$ that $S^{(i)}(x, \cdot)$ is irreducible. $S^{(i)}(x, \cdot)$ is also aperiodic because the proposal $q(x, y)$ has positive probability to leave the configuration x staying the same.

Since $S^{(i)}(x, \cdot)$ keeps π_i as the steady-state distribution, it finally follows from the standard Markov chain convergence theorem and the asymptotic equivalence (5) that $X^{(i)}$ is ergodic with π_i as its steady-state distribution. The proof is thus terminated. □

REMARK 3. The assumption (ii) is weaker than assuming that $T_{\mathrm{MH}}^{(i)}$ is irreducible for $i = 0, 1, \ldots, K - 1$, because we can see that essentially the function of the MH local move is to bridge adjacent energy sets, while the equi-energy jump allows jumps within an equi-energy set.

3.3. *Practical implementation.* There are some flexibilities in the practical implementation of the EE sampler. We provide some suggestions based on our own experience.

1. The choice of the temperature and energy ladder.
   Given the lowest and second highest energy levels H_0 and H_K, we found that setting the other energy levels by a geometric progression, or equivalently setting $\log(H_{i+1} - H_i)$ to be evenly spaced, often works quite well. The temperature could be chosen such that $(H_{i+1} - H_i)/T_i \approx c$, and we found that $c \in [1, 5]$ often works well.
2. The choice of K, the number of temperature and energy levels.
   The choice of K depends on the complexity of the problem. More chains and energy levels are usually needed if the target distribution is high-dimensional and multimodal. In our experience K could be roughly proportional to the dimensionality of the target distribution.
3. The equi-energy jump probability p_{ee}.
   In our experience taking $p_{\mathrm{ee}} \in [5\%, 30\%]$ often works quite well. See Section 3.4 for more discussion.

4. Self-adaptation of the MH-proposal step size.

   As the order i increases, the distribution π_i becomes more and more flat. Intuitively, to efficiently explore a flat distribution, one should use a large step size in the MH proposal, whereas for a rough distribution, the step size has to be small. Therefore, in the EE sampler each chain $X^{(i)}$ should have its own step size in the local MH exploration. In practice, however, it is often difficult to choose the right step sizes in the very beginning. One can hence let the sampler tune by itself the step sizes. For each chain, the sampler can from time to time monitor the acceptance rate in the MH local move, and increase (decrease) the step size by a fixed factor, if the acceptance rate is too high (low). Note that in this self-adaptation the energy-ring structure remains unchanged.
5. Adjusting the energy and temperature ladder.

   In many problems, finding a close lower bound H_0 for the energy function $h(x)$ is not very difficult. But in some cases, especially when $h(x)$ is difficult to optimize, one might find during the sampling that the energy value at some state is actually smaller than the pre-assumed lower bound H_0. If this happens, we need to adjust the energy ladder and the temperatures, because otherwise the energy sets D_j would not have the proper sizes to cover the state space, which could affect the sampling efficiency. The adjustment can be done by dynamically monitoring the sampler. Suppose after the ith chain $X^{(i)}$ is started, but before the $(i-1)$st chain gets started, we find that the lowest energy value $H_{\min}$ reached so far is smaller than H_0. Then we first reset $H_0 = H_{\min} - \beta$, where the constant $\beta > 0$, say $\beta = 2$. Next given H_i and the new H_0 we reset the in-between energy levels by a geometric progression, and if necessary add in more energy levels between H_0 and H_i (thus adding more chains) so that the sequence $H_{j+1} - H_j$ is still monotone increasing in j. The temperatures between $T_0 = 1$ and T_i are reset by $(H_{j+1} - H_j)/T_j \approx c$. With the energy ladder adjusted, the samples are regrouped to new energy rings. Note that since the chains $X^{(K)}, X^{(K-1)}, \ldots, X^{(i)}$ have already started, we do not change the values of $H_K, \ldots, H_i$ and $T_K, \ldots, T_i$, so that the target distributions $\pi_K, \ldots, \pi_i$ are not altered.

3.4. *A multimodal illustration.* As an illustration, we consider sampling from a two-dimensional normal mixture model taken from [23],

$$f(\mathbf{x}) = \sum_{i=1}^{20} \frac{w_i}{2\pi\sigma_i^2} \exp\left\{-\frac{1}{2\sigma_i^2}(\mathbf{x} - \boldsymbol{\mu}_i)'(\mathbf{x} - \boldsymbol{\mu}_i)\right\}, \tag{6}$$

where $\sigma_1 = \cdots = \sigma_{20} = 0.1$, $w_1 = \cdots = w_{20} = 0.05$, and the 20 mean vectors

$$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{20}) = \begin{pmatrix} 2.18 & 8.67 & 4.24 & 8.41 & 3.93 & 3.25 & 1.70 \\ 5.76' & 9.59' & 8.48' & 1.68' & 8.82' & 3.47' & 0.50' \\ \\ 4.59 & 6.91 & 6.87 & 5.41 & 2.70 & 4.98 & 1.14 \\ 5.60' & 5.81' & 5.40' & 2.65' & 7.88' & 3.70' & 2.39' \end{pmatrix}$$

$$\begin{matrix} 8.33 & 4.93 & 1.83 & 2.26 & 5.54 & 1.69 \\ 9.50\text{'} & 1.50\text{'} & 0.09\text{'} & 0.31\text{'} & 6.86\text{'} & 8.11 \end{matrix}\Bigg).$$

Since most local modes are more than 15 standard deviations away from the nearest ones [see Figure 2(a)], this mixture distribution poses a serious challenge for sampling algorithms, and thus serves as a good test. We applied the EE sampler to this problem. Since the minimum value of the energy function $h(\mathbf{x}) = -\log(f(\mathbf{x}))$ is around $-\log(\frac{5}{2\pi}) = 0.228$, we took $H_0 = 0.2$. K was set to 4, so only five chains were employed. The energy levels $H_1, \ldots, H_4$ were set by a geometric progression in the interval $[0, 200]$. The settings for energy levels and temperature ladders are tabulated in Table 1. The equi-energy jump probability p_{ee} was taken to be 0.1. The initial states of the chains $X^{(i)}$ were drawn uniformly from $[0, 1]^2$, a region far from the centers $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{20}$, so as to make the sampling challenging. The MH proposal is taken to be bivariate Gaussian: $X^{(i)}_{n+1} \sim N_2(X^{(i)}_n, \tau_i^2 I_2)$, where the initial MH proposal step size τ_i for the ith order chain $X^{(i)}$ was taken to be $0.25\sqrt{T_i}$. The step size was finely tuned later in the algorithm such that the acceptance ratio was in the range $(0.22, 0.32)$. After a burn-in period, each chain was run for 50,000 iterations. Figure 2 shows the samples generated in each chain: With the help of the higher-order chains, where the distributions are more flat, all the modes of the target distribution were successfully visited by $X^{(0)}$. The number of samples in each energy ring is reported in Table 1. One can see that for low-order chains the samples are mostly concentrated in the low-energy rings, while for high-order chains more samples are distributed in the high-energy rings.

As a comparison, we also applied parallel tempering (PT) [8] to this problem. The PT procedure also adopts a temperature ladder; it uses a swap between neighboring temperature chains to help the low-temperature chain move. We ran the PT to sample from (6) with the same parameter and initialization setting. The step size of PT was tuned to make the acceptance ratio of the MH move between 0.22 and 0.32. The exchange (swap) probability of PT was taken to be 0.1 to make it comparable with $p_{\text{ee}} = 0.1$ in the EE sampler, and in each PT exchange operation, $K = 4$ swaps were proposed to exchange samples in neighboring chains. The

TABLE 1
Sample size of each energy ring

	Energy rings				
Chain	**< 2.0**	**[2.0, 6.3)**	**[6.3, 20.0)**	**[20.0, 63.2)**	**≥ 63.2**
$X^{(0)}, T_0 = 1$	41631	8229	140	0	0
$X^{(1)}, T_1 = 2.8$	21118	23035	5797	50	0
$X^{(2)}, T_2 = 7.7$	7686	16285	22095	3914	20
$X^{(3)}, T_3 = 21.6$	3055	6470	17841	20597	2037
$X^{(4)}, T_4 = 60.0$	1300	2956	8638	20992	16114

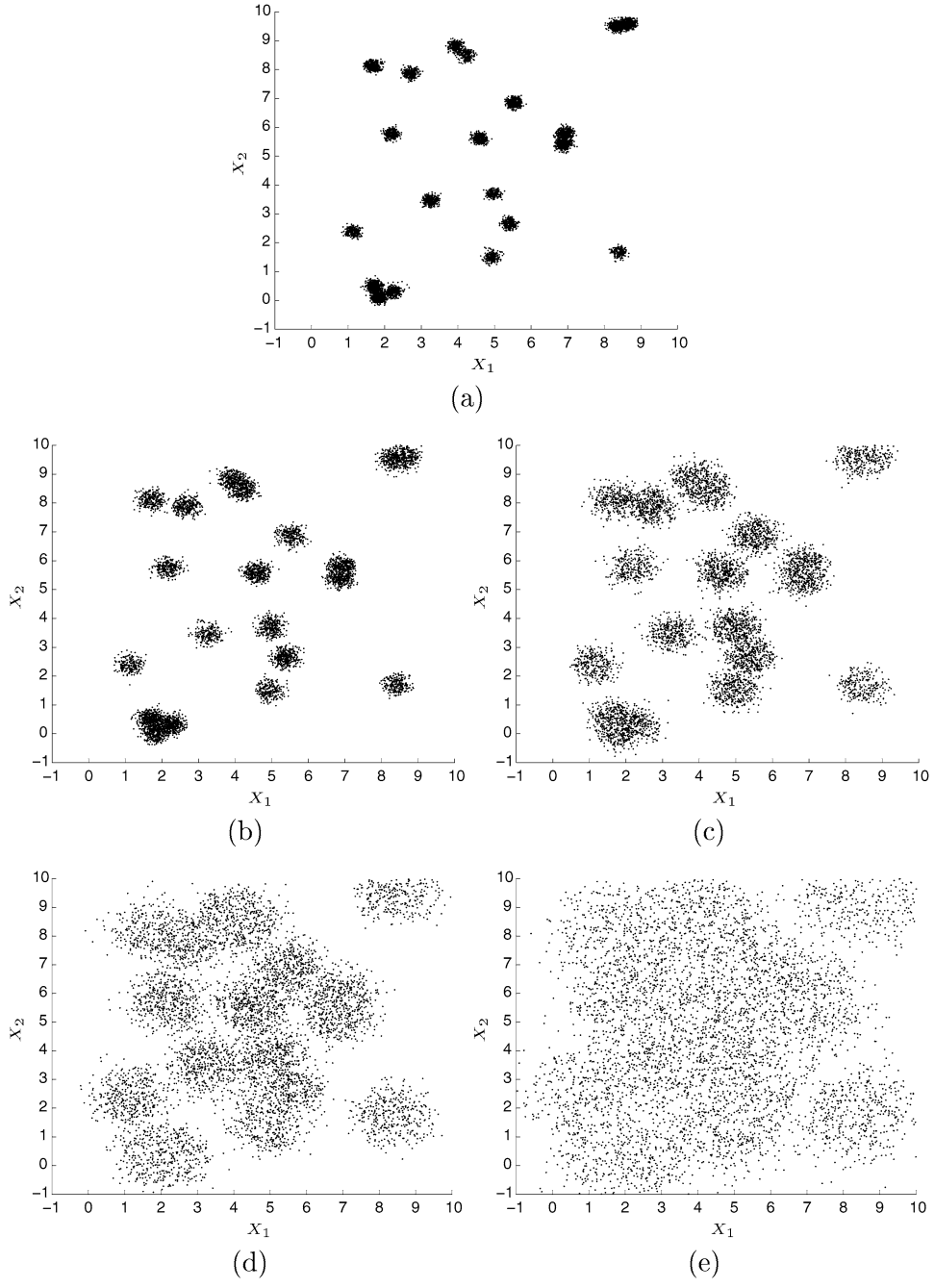

FIG. 2. *Samples generated from each chain of the EE sampler. The chains are sorted in ascending order of temperature and energy truncation.*

overall acceptance rates for the MH move in the EE sampler and parallel tempering were 0.27 and 0.29, respectively. In the EE sampler, the acceptance rate for the equi-energy jump was 0.82, while the acceptance rate for the exchange operation in PT was 0.59. Figure 3(a) shows the path of the last 2000 samples in $X^{(0)}$ for the EE sampler, which visited all the 20 components frequently; in comparison PT only visited 14 components in the same number of samples [Figure 3(b)]. As a further test, we ran the EE sampler and the PT 20 times independently, and sought to estimate the mean vector (EX_1, EX_2) and the second moment (EX_1^2, EX_2^2) using the samples generated from the target chain $X^{(0)}$. The results are shown in Table 2 (the upper half). It is clear that the EE sampler provided more accurate estimates with smaller mean squared errors.

To compare the mixing speed for the two sampling algorithms, we counted in each of the 20 runs how many times the samples visited each mode in the last 2000 iterations. We then calculated the absolute frequency error for each mode, $\text{err}_i = |\hat{f}_i - 0.05|$, where $\hat{f}_i$ is the sample frequency of the ith mode ($i = 1, \ldots, 20$) being visited. For each mode i, we calculated the median and the maximum of err_i over the 20 runs. Table 3 reports, for each mode, the ratio $R1$ of the median frequency error of PT over that of EE, and the ratio $R2$ of the maximum frequency error of PT over the corresponding value of EE. A two- to fourfold improvement by EE was observed. We also noted that EE did not miss a single mode in all runs, whereas PT missed some modes in each run. Table 3 also shows, out of the 20 runs, how many times each mode was missed by PT; for example, mode 1 was missed twice by PT over the 20 runs. The better global exploration ability of the EE sampler is thus evident. To further compare the two algorithms by their convergence speed, we tuned the temperature ladder for PT to achieve the *best* performance of PT; for example, we tuned the highest temperature T_4 of PT in the range $[5, 100]$. We observe that the best performance of PT such that it is not trapped by some local modes in 50,000 samples and that it achieves minimal sample autocorrelation is the setting associated with $T_4 = 10$. The sample autocorrelations of the target chain $X^{(0)}$ from the EE sampler and the optimal PT are shown in Figures 3(c) and (d), respectively; evidently the autocorrelation of the EE sampler decays much faster even compared with a well-tuned PT.

We also use this example to study the choice of p_{ee}. We took $p_{\text{ee}} = 0.1, 0.2, 0.3$ and 0.4, and ran the EE sampler 20 times independently for each value of p_{ee}. From the average mean squared errors for estimating (EX_1, EX_2) and (EX_1^2, EX_2^2) we find that the EE sampler behaves well when p_{ee} is between 0.1 and 0.3. When p_{ee} is increased to 0.4, the performance of the EE sampler worsens. In addition, we also noticed that the performance of the EE sampler is not sensitive to the value of p_{ee}, as long as it is in the range $[0.05, 0.3]$.

Next, we changed the weight and variance for each component in (6) such that $w_i \propto 1/d_i$ and $\sigma_i^2 = d_i/20$, where $d_i = \|\boldsymbol{\mu}_i - (5, 5)'\|$. The distributions closer to

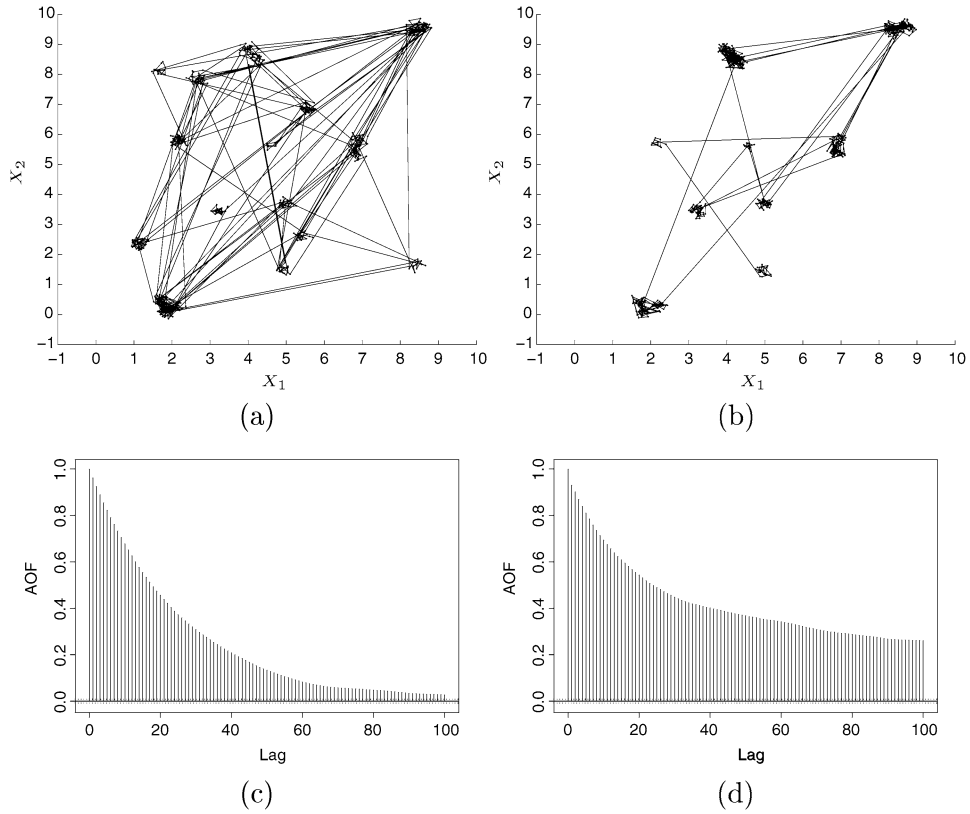

FIG. 3. *Mixture normal distribution with equal weights and variances. The sample path of the last* 2000 *iterations for* (a) *EE sampler and* (b) *parallel tempering. Autocorrelation plots for the samples from* (c) *EE sampler and* (d) *optimal parallel tempering.*

TABLE 2
Comparison of the EE sampler and PT for estimating the mixture normal distributions

	EX_1	EX_2	EX_1^2	EX_2^2
True value	4.478	4.905	25.605	33.920
EE	4.5019 (0.107)	4.9439 (0.139)	25.9241 (1.098)	34.4763 (1.373)
PT	4.4185 (0.170)	4.8790 (0.283)	24.9856 (1.713)	33.5966 (2.867)
MSE(PT)/MSE(EE)	2.7	3.8	2.6	3.8
True value	4.688	5.030	25.558	31.378
EE	4.699 (0.072)	5.037 (0.086)	25.693 (0.739)	31.433 (0.839)
PT	4.709 (0.116)	5.001 (0.134)	25.813 (1.122)	31.105 (1.186)
MSE(PT)/MSE(EE)	2.6	2.5	2.4	2.1

The numbers in parentheses are the standard deviations from 20 independent runs. The upper and bottom halves correspond to equal and unequal weights and variances, respectively.

TABLE 3
Comparison of mixing speed of the EE sampler and PT for the mixture normal distribution

	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}
$R1$	2.8	2.9	3.2	3.6	2.7	4.1	1.7	2.1	2.7	3.3
$R2$	1.6	4.5	2.5	4.9	2.5	2.1	2.2	2.3	2.3	2.5
PT_{mis}	2	3	1	5	4	0	2	2	3	1
	μ_{11}	μ_{12}	μ_{13}	μ_{14}	μ_{15}	μ_{16}	μ_{17}	μ_{18}	μ_{19}	μ_{20}
$R1$	1.3	2.5	2.0	4.4	2.2	2.5	2.0	1.9	2.6	1.4
$R2$	3.8	1.2	1.9	2.1	3.2	3.0	3.1	3.0	3.9	2.0
PT_{mis}	1	3	1	1	1	3	4	6	1	3

$R1$ is the ratio of median frequency error of PT over that of EE; $R2$ is the ratio of maximum frequency error of PT over that of EE. PT_{mis} reports the number of runs for which PT missed the individual modes. The EE sampler did not miss any of the 20 modes. All the statistics are calculated from the last 2000 samples in 20 independent runs.

(5, 5) have larger weights and lower energy. We used this example to test our strategy of dynamically updating the energy and temperature ladders (Section 3.3). We set the initial energy lower bound $H_0 = 3$, which is higher than the energy at any of the 20 modes (in practice, we could try to get a better initial value of H_0 by some local optimization). The highest energy level and temperature were set at 100 and 20, respectively. We started with five chains and dynamically added more chains if necessary. The strategy for automatically updating the proposal step size was applied as well. After drawing the samples we also calculated the first two sample moments from the target chain $X^{(0)}$ as simple estimates for the theoretical moments. The mean and standard deviation of the estimates based on 20 independent EE runs, each consisting of 10,000 iterations, are shown in Table 2 (the bottom half). The sample path for the last 1000 iterations and the autocorrelation plots are shown in Figure 4.

For comparison, PT was applied to this unequal weight case as well. PT used the same total number of chains as the EE sampler. The MH step size of PT was tuned to achieve the same acceptance rate. The temperature ladder of PT was also tuned so that the exchange operator in PT had the same acceptance rate as the equi-energy jump in EE, similarly to what we did in the previous comparison. With these well-tuned parameters, we ran PT for the same number of iterations and calculated the first two sample moments as we did for the EE samples. The results are reported in Table 2. It is seen that the EE sampler with the self-adaptation strategies provided more precise estimates (both smaller bias and smaller variance) in all the cases. Similar improvements in mixing speed and autocorrelation decay were also observed (Figure 4).

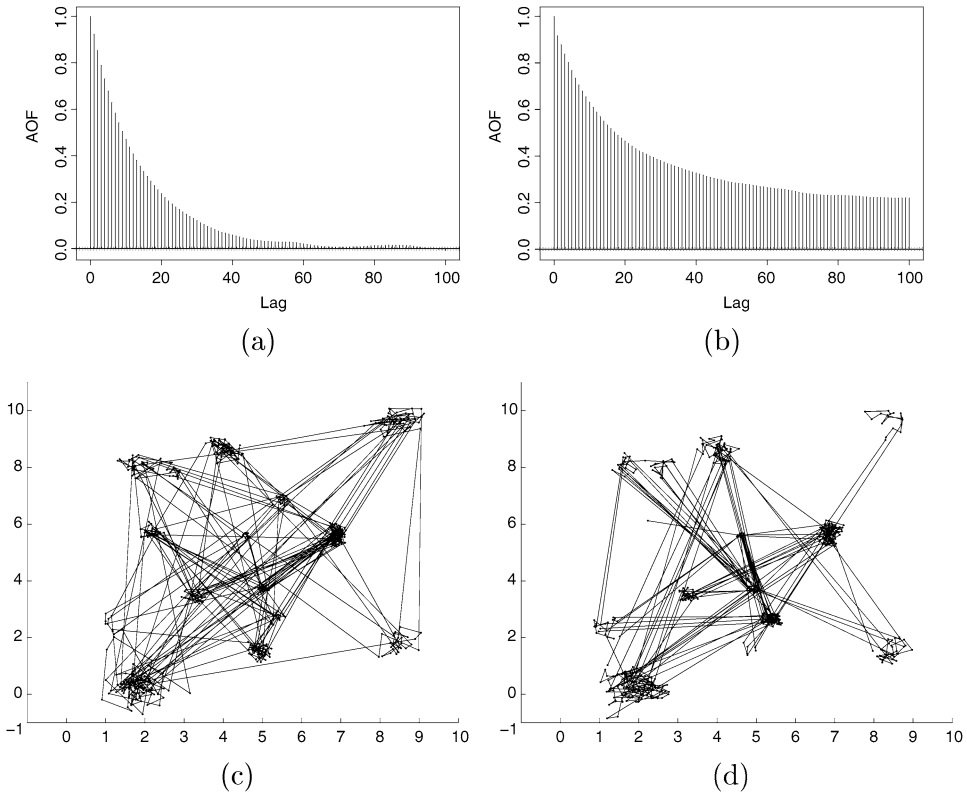

FIG. 4. *Mixture normal distribution with unequal weights and variances. Autocorrelation plots for the samples from* (a) *EE sampler and* (b) *parallel tempering. The sample path for the last* 1000 *iterations for* (c) *EE sampler and* (d) *parallel tempering.*

4. Calculating the density of states and the Boltzmann averages. The previous section illustrated the benefit of constructing the energy rings: It allows more efficient sampling through the equi-energy jump. By enabling one to look at the states within a given energy range, the energy rings also provide a direct means to study the microscopic structure of the state space on an energy-by-energy basis, that is, the microcanonical distribution. The EE sampler and the energy rings are thus well suited to study problems in statistical mechanics.

Starting from the Boltzmann distribution (1) one wants to study various aspects of the system. The density of states $\Omega(u)$, whose logarithm is referred to as the microcanonical entropy, plays an important role in the study, because in addition to the temperature–energy duality depicted in Section 2, many thermodynamic quantities can be directly calculated from the density of states, for example, the partition function $Z(T)$, the internal energy $U(T)$, the specific heat $C(T)$ and the

free energy $F(T)$ can be calculated via

$$Z(T) = \int \Omega(u) e^{-u/T}\, du,$$

$$U(T) = \frac{\int u\, \Omega(u) e^{-u/T}\, du}{\int \Omega(u) e^{-u/T}\, du},$$

$$C(T) = \frac{\partial U(T)}{\partial T} = \frac{1}{T^2}\left[\frac{\int u^2 \Omega(u) e^{-u/T}\, du}{\int \Omega(u) e^{-u/T}\, du} - \left(\frac{\int u \Omega(u) e^{-u/T}\, du}{\int \Omega(u) e^{-u/T}\, du}\right)^2\right],$$

$$F(T) = -T \log(Z(T)).$$

Since the construction of the energy rings is an integral part of the EE sampler, it leads to a simple way to estimate the density of states. Suppose we have a discrete system to study, and after performing the EE sampling on the distributions

$$\pi_i(x) \propto \exp(-h_i(x)), \qquad h_i(x) = \frac{1}{T_i}\big(h(x) \vee H_i\big), \qquad 0 \le i \le K,$$

we obtain the energy rings $\hat{D}_j^{(i)}$ $(0 \le i, j \le K)$, each $\hat{D}_j^{(i)}$ corresponding to an a priori energy range $[H_j, H_{j+1})$. To calculate the density of states for the discrete system we can further divide the energy rings into subsets such that each subset corresponds to one energy level. Let m_{iu} denote the total number of samples in the ith chain $X^{(i)}$ that have energy u. Clearly $\sum_{u \in [H_j, H_{j+1})} m_{iu} = |\hat{D}_j^{(i)}|$. Under the distribution π_i

$$P_{\pi_i}\big(h(X) = u\big) = \frac{\Omega(u) e^{-(u \vee H_i)/T_i}}{\sum_v \Omega(v) e^{-(v \vee H_i)/T_i}}. \tag{7}$$

Since the density of states $\Omega(u)$ is common for each π_i, we can combine the sample chains $X^{(i)}$, $0 \le i \le K$, to estimate $\Omega(u)$. Denote $m_{i\bullet} = \sum_u m_{iu}$, $m_{\bullet u} = \sum_i m_{iu}$ and $a_{iu} = e^{-(u \vee H_i)/T_i}$ for notational ease. Pretending we have independent multinomial observations,

$$(\ldots, m_{iu}, \ldots) \sim \text{multinomial}\left(m_{i\bullet}; \ldots, \frac{\Omega(u) a_{iu}}{\sum_v \Omega(v) a_{iv}}, \ldots\right), \qquad i = 0, 1, \ldots, K, \tag{8}$$

the MLE of $\Omega(u)$ is then

$$\hat{\Omega} = \arg\max_{\Omega} \left\{ \sum_u m_{\bullet u} \log(\Omega(u)) - \sum_i m_{i\bullet} \log\left(\sum_v \Omega(v) a_{iv}\right)\right\}. \tag{9}$$

Since $\Omega(u)$ is specified up to a scale change [see (7)], to estimate the relative value we can without loss of generality set $\Omega(u_0) = 1$ for some u_0. The first-order condition of (9) gives

$$\frac{m_{\bullet u}}{\hat{\Omega}(u)} - \sum_i \frac{m_{i\bullet} a_{iu}}{\sum_v \hat{\Omega}(v) a_{iv}} = 0 \qquad \text{for all } u, \tag{10}$$

which can be used to compute $\hat{\Omega}(u)$ through a simple iteration,

$$\hat{\Omega}(u) = m_{\bullet u} \Big/ \sum_i \frac{m_{i\bullet} a_{iu}}{\sum_v \hat{\Omega}(v) a_{iv}}. \tag{11}$$

A careful reader might question the independent multinomial assumption (9). But it is only used to motivate (10), which itself can be viewed as a moment equation and is valid irrespective of the multinomial assumption.

With the density of states estimated, suppose one wants to investigate how the Boltzmann average $E(g(X); T) = \frac{\sum_x g(x)\exp(-h(x)/T)}{\sum_x \exp(-h(x)/T)}$ varies as a function of temperature T (e.g., phase transition). Then we can write (see Lemma 1)

$$E\big(g(X); T\big) = \frac{\sum_u \Omega(u) e^{-u/T} v_g(u)}{\sum_u \Omega(u) e^{-u/T}}.$$

To estimate the microcanonical average $v_g(u) = E(g(X)|h(X) = u)$, we can simply calculate the sample average over the energy slice $\{x : h(x) = u\}$ for each chain $X^{(0)}, \ldots, X^{(K)}$ and combine them using weights proportional to the energy slice sample size m_{iu} for $i = 0, 1, \ldots, K$.

So far we have focused on discrete systems. In continuous systems to estimate $\Omega(u)$ and $v_g(u) = E(g(X)|h(X) = u)$ we can simply discretize the energy space with an acceptable resolution and follow the preceding approach to use the EE sampler and the energy rings.

To illustrate our method to calculate the density of states and the Boltzmann averages, let us consider a multidimensional normal distribution $f(\mathbf{x}; T) = \frac{1}{Z(T)} \exp[-h(\mathbf{x})/T]$ with temperature T and the energy function $h(\mathbf{x}) = \frac{1}{2}\sum_i^n x_i^2$. This corresponds to a system of n uncoupled harmonic oscillators. Since the analytical result is known in this case, we can check our numerical calculation with the exact values. Suppose we are interested in estimating $E(X_1^2; T)$ and the ratio of normalizing constants $Z(T)/Z(1)$ for T in $[1, 5]$. The theoretical density of states (from the χ^2 result in this case) is $\Omega(u) \propto u^{n/2-1}$ and the microcanonical average $v_g(u) = E(X_1^2|h(\mathbf{X}) = u) = 2u/n$. Our goal is to estimate $\Omega(u)$ and $v_g(u)$ and then combine them to estimate both $E(X_1^2; T)$ and $Z(T)/Z(1)$ as functions of the temperature T.

We took $n = 4$ and applied the EE sampler with five chains to sample from the four-dimensional distribution. The energy levels $H_0, H_1, \ldots, H_4$ were assigned by a geometric progression along the interval $[0, 50]$. The temperatures were accordingly set between 1 and 20. The equi-energy jump probability p_{ee} was taken to be 0.05. Each chain was run for 150,000 iterations (the first 50,000 iterations were the burn-in period); the sampling was repeated ten times to calculate the standard error of our estimates. Since the underlying distribution is continuous, to estimate $\Omega(u)$ and $v_g(u)$ each energy ring was further evenly divided into 20 small intervals. The estimates $\hat{\Omega}(u)$ and $\hat{v}_g(u)$ were calculated from the recursion (11) and the combined energy slice sample average, respectively. Figure 5(a) and (b) shows

$\hat{\Omega}(u)$ and $\hat{v}_g(u)$ compared with their theoretical values. Using $\hat{\Omega}(u)$ and $\hat{v}_g(u)$, we estimated $E(X_1^2; T)$ and $Z(T)/Z(1)$ by

$$E(X_1^2; T) \approx \frac{\sum_u \hat{v}_g(u)\hat{\Omega}(u)e^{-u/T}\,\Delta u}{\sum_u \hat{\Omega}(u)e^{-u/T}\,\Delta u},$$

$$\frac{Z(T)}{Z(1)} \approx \frac{\sum_u \hat{\Omega}(u)e^{-u/T}\,\Delta u}{\sum_u \hat{\Omega}(u)e^{-u}\,\Delta u}.$$

Letting T vary in $[1, 5]$ results in the estimated curves in Figure 5(c) and (d). One can see from the figure that our estimations are very precise compared to the theoretical values. In addition, our method has the advantage that we are able to

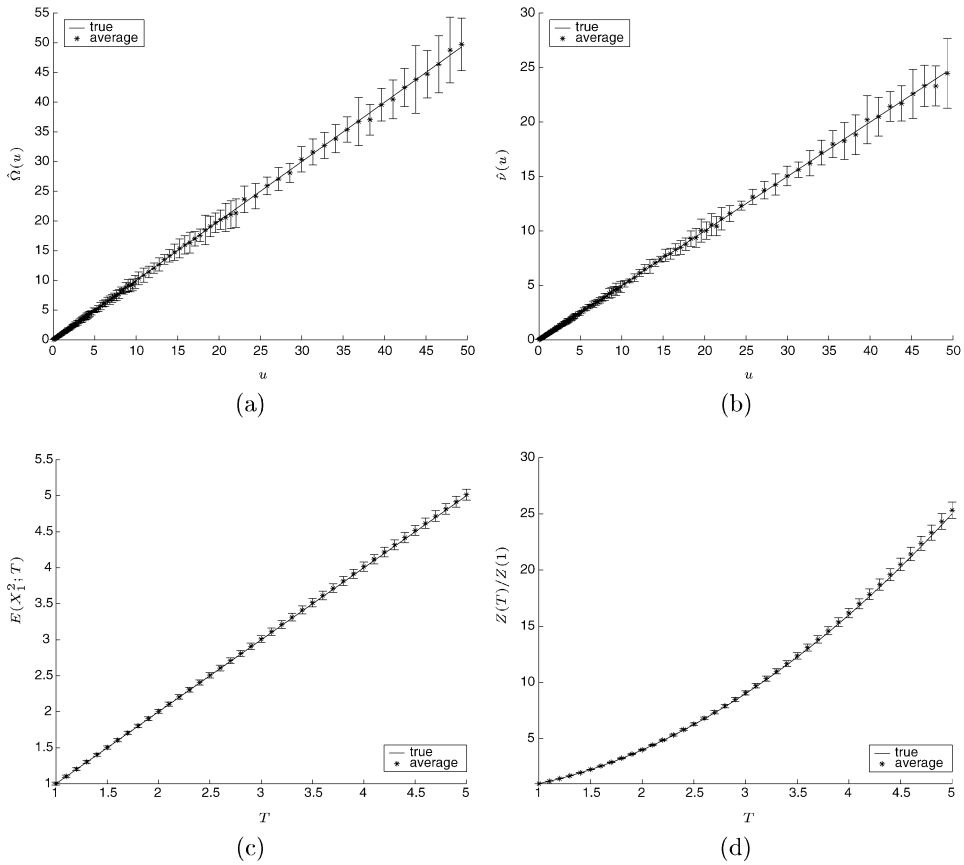

FIG. 5. *The density of states calculation for the four-dimensional normal distribution.* (a) *Estimated density of states* $\hat{\Omega}(u)$; (b) *estimated micro-canonical average* $\hat{v}(u)$; (c) *estimated* $E(X_1^2; T)$ *and* (d) *estimated* $Z(T)/Z(1)$ *for* T *varying from* 1 *to* 5. *The error bars represent plus or minus twice the standard deviation from ten independent runs.*

construct estimates for a wide temperature range using one simulation that involves only five temperatures.

We further tested our method on a four-dimensional normal mixture distribution with the energy function

$$h(\mathbf{x}) = -\log[\exp(-\|\mathbf{x}-\boldsymbol{\mu}_1\|^2) + 0.25\exp(-\|\mathbf{x}-\boldsymbol{\mu}_2\|^2)], \tag{12}$$

where $\boldsymbol{\mu}_1 = (3,0,0,0)'$ and $\boldsymbol{\mu}_2 = (-3,0,0,0)'$. For the Boltzmann distribution $\frac{1}{Z(T)}\exp(-h(\mathbf{x})/T)$, we are interested in estimating the probability $P(X_1 > 0; T)$ and studying how it varies as T changes. It is easy to see if $T = 1$ this probability will be $0.8 = 1/1.25$, the mixture proportion, and it decreases as T becomes larger. We applied the EE sampler to this problem with the same parameter setting as in the preceding 4D normal example. Using the energy rings, we calculated the estimates $\hat{\Omega}(u)$ and $\hat{\nu}_g(u)$ and then combined them to estimate $P(X_1 > 0; T)$. Figure 6 plots the estimates. It is interesting to note from the figure that the density of states for the mixture model has a change point at energy $u = 1.4$. This is due to the fact that the energy at the mode $\boldsymbol{\mu}_2$ is about $1.4(\approx -\log 0.25)$, and hence for $u < 1.4$ all the samples are from the first mode $\boldsymbol{\mu}_1$, and for $u > 1.4$ the samples come from both modes, whence a change point appears. A similar phenomenon occurs in the plot for $\nu_g(u) = P(X_1 > 0 | h(\mathbf{X}) = u)$. The combined estimate of $P(X_1 > 0; T)$ in Figure 6 was checked and agreed very well with the exact values from numerical integration.

5. Statistical estimation with the EE sampler. In statistical inference, usually after obtaining the Monte Carlo samples the next goal is to estimate some statistical quantities. Unlike statistical mechanical considerations, in statistical inference problems often one is only interested in one target distribution (i.e., only one temperature). Suppose the expected value $E_{\pi_0} g(X)$ under the target distribution $\pi_0 = \pi$ is of interest. A simple estimate is the sample mean under the chain $X^{(0)}$ ($T = 1$). This, however, does not use the samples optimally in that it essentially throws away all the other sampling chains $X^{(1)}, \ldots, X^{(K)}$. With the help of the energy rings, in fact all the chains can be combined together to provide a more efficient estimate. One way is to use the energy-temperature duality as discussed above. Here we present an alternative, more direct method: We directly work with the (finite number of) expectations within each energy set. For a continuous problem, this allows us to avoid dealing with infinitesimal energy intervals, which are needed in the calculation of microcanonical averages.

The starting point is the identity

$$E_{\pi_0} g(X) = \sum_{j=0}^{K} p_j E_{\pi_0}\{g(X) | X \in D_j\}, \tag{13}$$

where $p_j = P_{\pi_0}(X \in D_j)$, which suggests that we can first estimate p_j and $G_j = E_{\pi_0}\{g(X)|X \in D_j\}$, and then conjoin them. A naive estimate of G_j is the sample

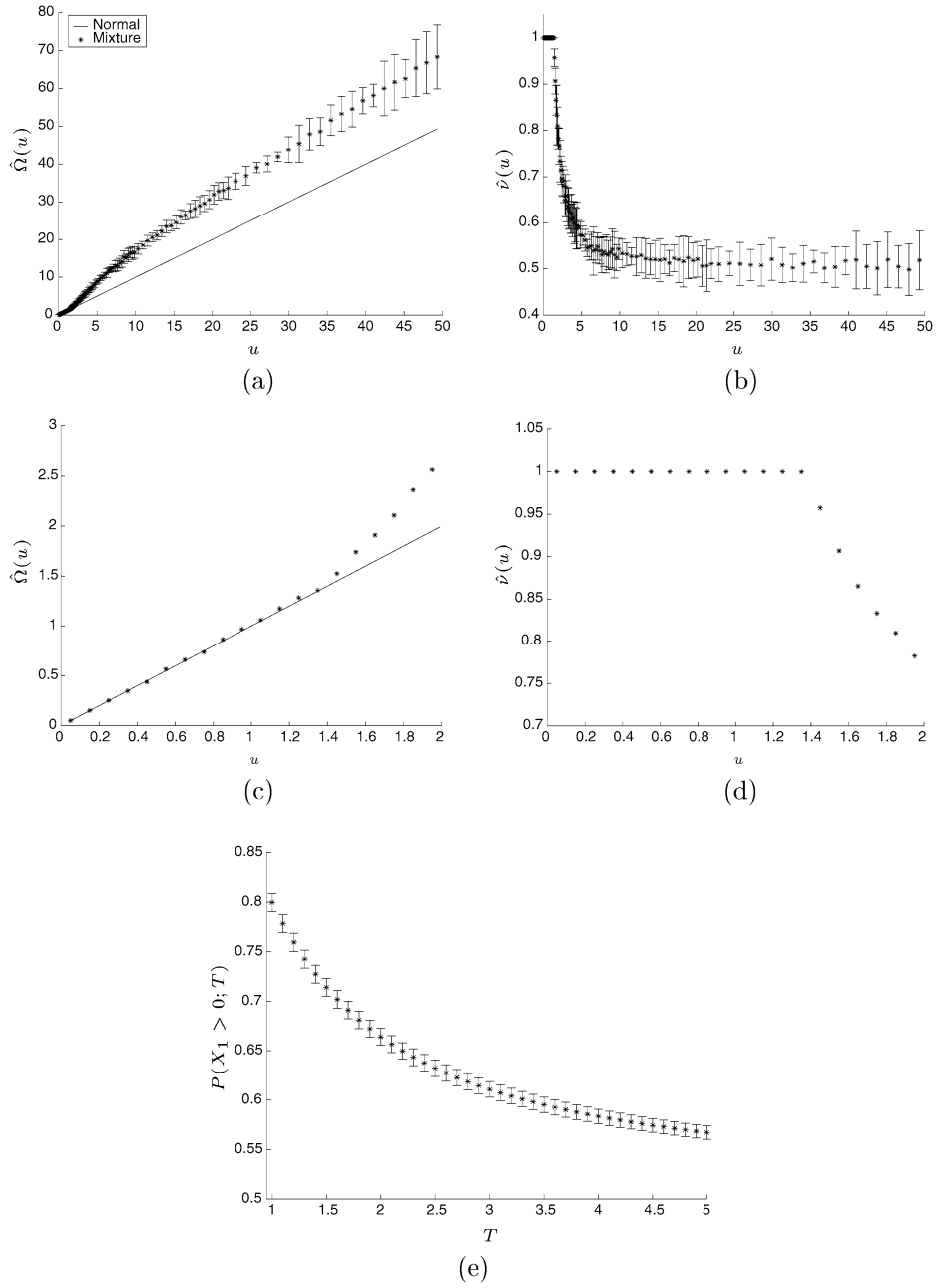

FIG. 6. *Four-dimensional normal mixture distribution*. (a) *Estimated density of states* $\hat{\Omega}(u)$; (b) *estimated microcanonical average* $\hat{\nu}(u)$; (c) *detailed plot for* $\hat{\Omega}(u)$ *around the change point*; (d) *detailed plot for* $\hat{\nu}(u)$ *around the change point*; (e) *estimated* $P(X_1 > 0; T)$ *for* T *in* [1, 5]. *For comparison, the theoretical* $\Omega(u)$ *from the* 4*D normal example is also shown in* (a) *and* (c). *The error bars represent plus or minus twice the standard deviation from ten independent runs.*

average of $g(x)$ within the energy ring $\hat{D}_j^{(0)}$. But a better way is to use the energy rings $\hat{D}_j^{(i)}$ for $i = 0, 1, \ldots, K$ together, because for each i the importance-weighted

$$\hat{G}_j^{(i)} = \frac{\sum_{X \in \hat{D}_j^{(i)}} g(X) w^{(i)}(X)}{\sum_{X \in \hat{D}_j^{(i)}} w^{(i)}(X)},$$

where $w^{(i)}(x) = \exp\{h_i(x) - h(x)\}$, is a consistent estimate of G_j. We can thus put proper weights on them to combine them.

Since it is quite often that a number of different estimands g are of interest, a conceptually simple and effective method is to weight $\hat{G}_j^{(i)}$, $i = 0, 1, \ldots, K$, proportionally to their *effective sample sizes* [15],

$$ESS_j^{(i)} = \frac{|\hat{D}_j^{(i)}|}{1 + \mathrm{Var}_{\pi_i}\{w^{(i)}(X) | X \in D_j\}/(E_{\pi_i}\{w^{(i)}(X) | X \in D_j\})^2},$$

which measures the effectiveness of a given importance-sampling scheme, and does not involve the specific estimand. $ESS_j^{(i)}$ can be easily estimated by using the sample mean and variance of $w^{(i)}(X)$ for X in the energy ring $\hat{D}_j^{(i)}$.

To estimate the energy-ring probability $p_j = P_{\pi_0}(X \in D_j)$, one can use the sample size proportion $\frac{|\hat{D}_j^{(0)}|}{\sum_{k=0}^{K} |\hat{D}_k^{(0)}|}$. But again it is better to use all the chains, because for any i,

$$\hat{p}_j^{(i)} = \frac{\sum_{X \in \hat{D}_j^{(i)}} w^{(i)}(X)}{\sum_{X \in \hat{D}^{(i)}} w^{(i)}(X)}, \qquad \text{where } \hat{D}^{(i)} = \bigcup_j \hat{D}_j^{(i)}$$

is a consistent estimate of p_j. To combine them properly, we can weight them inversely proportional to their variances. A simple delta method calculation (disregarding the dependence) gives the asymptotic variance of $\hat{p}_j^{(i)}$,

$$V_j^{(i)} = \frac{1}{n_i} \frac{E_{\pi_i}[(I(X \in D_j) - p_j) w^{(i)}(X)]^2}{[E_{\pi_i} w^{(i)}(X)]^2},$$

where n_i is the total number of samples under the ith chain $X^{(i)}$. $V_j^{(i)}$ can be estimated by

$$\begin{aligned}\hat{V}_j^{(i)} &= \frac{\sum_{X \in \hat{D}^{(i)}} [(I(X \in D_j) - \tilde{p}_j) w^{(i)}(X)]^2}{(\sum_{X \in \hat{D}^{(i)}} w^{(i)}(X))^2} \\ &= (1 - 2\tilde{p}_j) \frac{\sum_{X \in \hat{D}_j^{(i)}} [w^{(i)}(X)]^2}{(\sum_{X \in \hat{D}^{(i)}} w^{(i)}(X))^2} + \tilde{p}_j^2 \frac{\sum_{X \in \hat{D}^{(i)}} [w^{(i)}(X)]^2}{(\sum_{X \in \hat{D}^{(i)}} w^{(i)}(X))^2},\end{aligned}$$

where $\tilde{p}_j$ is a consistent estimate of p_j. Since $\hat{V}_j^{(i)}$ requires an a priori $\tilde{p}_j$, a simple iterative procedure can be applied to obtain a good estimate of p_j, starting from the naive $\hat{p}_j^{(0)}$. In practice to ensure the numerical stability of the variance estimate $\hat{V}_j^{(i)}$ it is recommended that $\hat{p}_j^{(i)}$ be included in the combined estimate only if the sample size $|\hat{D}_j^{(i)}|$ is reasonably large, say more than 50. When we get our estimates for p_j, we need to normalize them before plugging in (13) to calculate our combined estimates.

Since our energy-ring estimate of $E_{\pi_0} g(X)$ employs all the sampling chains, one expects it to increase the estimation efficiency by a large margin compared with the naive method.

The two-dimensional normal mixture model (*continued*). To illustrate our estimation strategy, consider again the 2D mixture model (6) in Section 3.4 with equal weights and variances $\sigma_1 = \cdots = \sigma_{20} = 0.1$, $w_1 = \cdots = w_{20} = 0.05$.

Suppose we want to estimate the functions EX_1^2, EX_2^2, Ee^{-10X_1} and Ee^{-10X_2} and the tail probabilities

$$p_1 = P\big(X_1 > 8.41, X_2 < 1.68 \text{ and } \sqrt{(X_1 - 8.41)^2 + (X_2 - 1.68)^2} > 4\sigma\big),$$

$$p_2 = P(X_1^2 + X_2^2 > 175).$$

After obtaining the EE samples (as described in Section 3.4), we calculated our energy-ring estimates. For comparison, we also calculated the naive estimates based on the target chain $X^{(0)}$ only. The calculation was repeated for the 20 independent runs of the EE sampler under the same parameter settings. Table 4 reports the result. Evidently, the energy-ring estimates with both smaller bias *and* smaller variance are much more precise than the naive estimates. The mean squared error has been reduced by at least 28% in all cases. The improvement over the naive

TABLE 4
Comparison of the energy-ring estimates with the naive estimates that use $X^{(0)}$ only

	EX_1^2	EX_2^2	Ee^{-10X_1}	Ee^{-10X_2}	p_1	p_2
True value	**25.605**	**33.920**	**9.3e-7**	**0.0378**	**4.2e-6**	**6.7e-5**
Energy ring estimates	25.8968 (0.9153)	34.2902 (1.1579)	8.8e-7 (1.2e-7)	0.0379 (0.0044)	4.5e-6 (1.5e-6)	6.4e-5 (2.0e-5)
Naive estimates	25.9241 (1.0982)	34.4763 (1.3733)	8.7e-7 (1.5e-7)	0.0380 (0.0052)	1.0e-5 (2.5e-5)	7.3e-5 (6.2e-5)
MSER	71%	67%	57%	72%	0.34%	11%

The numbers in parentheses are the standard deviations from 20 independent runs. MSER is defined as MSE_1/MSE_2, where MSE_1 and MSE_2 are the mean squared errors of the energy-ring estimates and the naive estimates, respectively.

method is particularly dramatic in the tail probability estimation, where the MSEs experienced a more than ninefold reduction.

6. Applications of the EE sampler. We will apply the EE sampler and the estimation strategy to a variety of problems in this section to illustrate its effectiveness. The first example is a mixture regression problem. The second example involves motif discovery in computational biology. In the third one we study the thermodynamic property of a protein folding model.

6.1. *Mixture exponential regression.* Suppose we observe data pairs $(\mathbf{Y}, \mathbf{X}) = \{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$ from the mixture model

$$y_i \sim \begin{cases} \text{Exp}[\theta_1(\mathbf{x}_i)], & \text{with probability } \alpha, \\ \text{Exp}[\theta_2(\mathbf{x}_i)], & \text{with probability } 1-\alpha, \end{cases} \tag{14}$$

where $\theta_j(\mathbf{x}_i) = \exp[\boldsymbol{\beta}_j^T \mathbf{x}_i]$ $(j = 1, 2)$, $\text{Exp}(\theta)$ denotes an exponential distribution with mean θ and $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ are p-dimensional vectors. Given the covariates $\mathbf{x}_i$ and the response variable y_i, one wants to infer the regression coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and the mixture probability α. The likelihood of the observed data is

$$\begin{aligned} &P(\mathbf{Y}, \mathbf{X}|\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \\ &\quad \propto \prod_{i=1}^{n} \left[\frac{\alpha}{\theta_1(\mathbf{x}_i)} \exp\left(-\frac{y_i}{\theta_1(\mathbf{x}_i)}\right) + \frac{1-\alpha}{\theta_2(\mathbf{x}_i)} \exp\left(-\frac{y_i}{\theta_2(\mathbf{x}_i)}\right) \right]. \end{aligned} \tag{15}$$

If we put a Beta(1, 1) prior on α, and a multivariate normal $N(\mathbf{0}, \sigma^2 \mathbf{I})$ on $\boldsymbol{\beta}_j$ $(j = 1, 2)$, the energy function, defined as the negative log-posterior density, is

$$\begin{aligned} h(\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &= -\log P(\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\mathbf{Y}, \mathbf{X}) \\ &= -\log P(\mathbf{Y}, \mathbf{X}|\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + \frac{1}{2\sigma^2} \sum_{k=1}^{2} \sum_{j=1}^{p} \boldsymbol{\beta}_{kj}^2 + C, \end{aligned} \tag{16}$$

where C is a constant. Since $h(\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = h(1-\alpha, \boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$ (i.e., nonidentifiable), the posterior distribution has multiple modes in the $(2p+1)$-dimensional parameter space. This example thus serves as a good model to test the performance of the EE sampler in high-dimensional multimodal problems.

We simulated 200 data pairs with the following parameter setting: $\alpha = 0.3$, $\boldsymbol{\beta}_1 = (1, 2)'$, $\boldsymbol{\beta}_2 = (4, 5)'$ and $\mathbf{x}_i = (1, u_i)'$, with the u_i's independently drawn from Unif(0, 2). We took $\sigma^2 = 100$ in the prior normal distributions for the regression coefficients. After a local minimization of the energy function (16) from several randomly chosen states in the parameter space, the minimum energy $H_{\min}$ is found to be around $H_{\min} \approx -1740.8$ (this value is not crucial in the EE sampler, since it can adaptively adjust the energy and temperature ladder; see Section 3.3).

We then applied the EE sampler with eight chains to sample from the posterior distribution. The energy ladder was set between $H_{\min}$ and $H_{\min} + 100$ in a geometric progression, and the temperatures were between 1 and 30. The equi-energy jump probability p_{ee} was taken to be 0.1. Each chain was run for 15,000 iterations with the first 5000 iterations serving as a burn-in period. The overall acceptance rates for the MH move and the equi-energy jump were 0.27 and 0.80, respectively. Figure 7(a) to (c) shows the sample marginal posterior distributions of α, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ from the target chain $X^{(0)}$. It is clear that the EE sampler visited the two modes, equally frequently in spite of the long distance between the two modes, and the samples around each mode were centered at the true parameters. Furthermore, since the posterior distribution for this problem has two symmetric modes in the parameter space due to the nonidentifiability, by visiting the two modes equally frequently the EE sampler demonstrates its capability of global exploration (as opposed to being trapped by a local mode).

For comparison, we also applied PT with eight chains to this problem under the same parameter setting, where the acceptance rates for the MH move and the exchange operator were 0.23 and 0.53, respectively. To compare their ability to escape local traps, we calculated the frequency with which the samples stayed at one particular mode in the lowest temperature chain ($T_0 = 1$). This frequency was 0.55 for the EE samplers and 0.89 for PT, indicating that the EE sampler visited the two modes symmetrically, while PT tended to be trapped at one mode for a long time. We further tuned the temperature ladder for PT with the highest temperature varying from 10 to 50. For each value of the highest temperature, the temperature ladder was set to decrease with a geometric rate. We observed that the sample autocorrelations decrease with the decrease of the highest temperature, that is, the denser the temperature ladder, the smaller the sample autocorrelation. However, the tradeoff is that with lower temperature, PT tends to get trapped in one local mode. For instance, PT was totally trapped to one particular mode if we set the highest temperature to be 10 although with this temperature ladder PT showed autocorrelation comparable with that of EE. On the other hand, if we increased the temperatures, PT would be able to jump between the two modes, but the autocorrelation also increased since the exchange rates became lower. But even with the highest temperature raised to 50, 80% of the PT samples were trapped to one mode and the posterior distributions were severely asymmetric. The autocorrelation plots for β_{11} are compared in Figure 7(d) and (e), where one sees that autocorrelation of the EE sampler decays faster than that of PT. We also plot the autocorrelation for PT with highest temperature $T_7 = 10$ in Figure 7(f), which corresponded to the minimal autocorrelation reached by PT among our tuning values, but the local trapping of PT with this parameter setting is severely pronounced.

6.2. *Motif sampling in biological sequences.* A central problem in biology is to understand how gene expression is regulated in different cellular processes. One

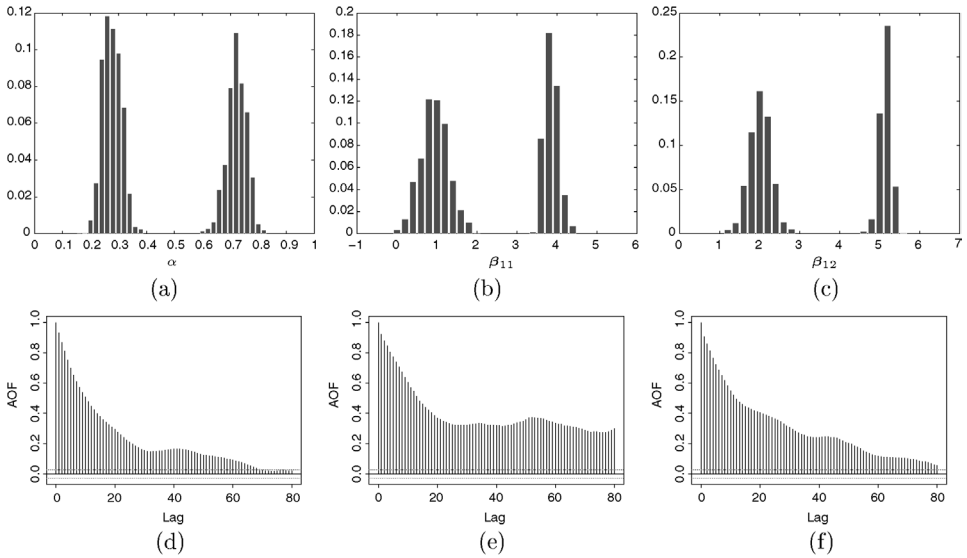

FIG. 7. *Statistical inference for the mixture exponential regression model. Marginal posterior distribution for* (a) α; (b) β_{11}; *and* (c) β_{12} (*the marginal distributions for* $\boldsymbol{\beta}_2$ *are similar to those for* $\boldsymbol{\beta}_1$ *and thus are not plotted here*). *Autocorrelation plot for the samples from* (d) *EE sampler*; (e) *PT with* $T_7 = 30$; *and* (f) *PT with* $T_7 = 10$.

important means for gene regulation is through the interaction between transcription factors (TF's) and their binding sites (viz., the sites that the TF recognizes and binds to in the regulatory regions of the gene). The common pattern of the recognition sites of a TF is called a binding motif, whose identification is the first step for understanding gene regulation. It is very time-consuming to determine these binding sites experimentally, and thus computational methods have been developed in the past two decades for discovering novel motif patterns and TF binding sites.

Some early methods based on site consensus used a progressive alignment procedure [36] to find motifs. A formal statistical model for the position-specific weight matrix (PWM)-based method was described in [21] and a complete Bayesian method was given in [25]. Based on a missing data formulation, the EM algorithm [1, 21] and the Gibbs sampler [20] were employed for motif discovery. The model has been generalized to discover modules of several clustered motifs simultaneously via data augmentation with dynamic programming [41]. See [14] for a recent review.

From a sampling point of view motif discovery poses a great challenge, because it is essentially a combinatorial problem, which makes the samplers very vulnerable to being trapped in numerous local modes. We apply the EE sampler to this problem under a complete Bayesian formulation to test its capability.

6.2.1. *Bayesian formulation of motif sampling.* The goal of motif discovery is to find the binding sites of a common TF (i.e., positions in the sequences

that correspond to a common pattern). Let $\mathbf{S}$ denote a set of M (promoter region) sequences, each sequence being a string of four nucleotides, A, C, G or T. The lengths of the sequences are $L_1, L_2, \ldots,$ and L_M. For notational ease, let $\mathbf{A} = \{A_{ij}, i = 1, 2, \ldots, M, j = 1, 2, \ldots, L_i\}$ be the indicator array such that $A_{ij} = 1$ if the jth position on the ith sequence is the starting point for a motif site and $A_{ij} = 0$ otherwise. In the motif sampling problem, a first-order Markov chain is used to model the background sequences; its parameters $\boldsymbol{\theta}_0$ (i.e., the transition probabilities) are estimated from the entire sequence data prior to the motif search. We thus effectively assume that $\boldsymbol{\theta}_0$ is known a priori. Given $\mathbf{A}$, we further denote the aligned motif sites by $\mathbf{S}(\mathbf{A})$, the nonsite background sequences by $\mathbf{S}(\mathbf{A}^c)$ and the total number of motif sites by $|\mathbf{A}|$. The motif width w is treated as known. The common pattern of the motif is modeled by a product multinomial distribution $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_w)$, where each $\boldsymbol{\theta}_i$ is a probability vector of length 4 for the preferences of the nucleotides (A, C, G, T) in the motif column i. See Figure 8 for an illustration of the motif model.

To write down the joint distribution of the complete data and all the parameters, it is assumed a priori that a randomly selected segment of width w has probability p_0 to be a motif site (p_0 is called the "site abundance" parameter). The joint distribution function has the form

$$
\begin{aligned}
P(\mathbf{S}, \mathbf{A}, \boldsymbol{\Theta}, p_0) &= P(\mathbf{S}|\mathbf{A}, \boldsymbol{\Theta}) P(\mathbf{A}|p_0) \pi(\boldsymbol{\Theta}) \pi(p_0) \\
&= \frac{P(\mathbf{S}(\mathbf{A})|\mathbf{A}, \boldsymbol{\Theta})}{P(\mathbf{S}(\mathbf{A})|\mathbf{A}, \boldsymbol{\theta}_0)} P(\mathbf{S}|\boldsymbol{\theta}_0) p_0^{|\mathbf{A}|} (1 - p_0)^{L - |\mathbf{A}|} \pi(\boldsymbol{\Theta}) \pi(p_0) \qquad (17) \\
&\propto \frac{1}{P(\mathbf{S}(\mathbf{A})|\mathbf{A}, \boldsymbol{\theta}_0)} \prod_{i=1}^{w} \boldsymbol{\theta}_i^{\mathbf{c}_i + \boldsymbol{\beta}_i - 1} p_0^{|\mathbf{A}| + a - 1} (1 - p_0)^{L - |\mathbf{A}| + b - 1},
\end{aligned}
$$

where we put a product Dirichlet prior $\pi(\boldsymbol{\Theta})$ with parameter $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_w)$ on $\boldsymbol{\Theta}$

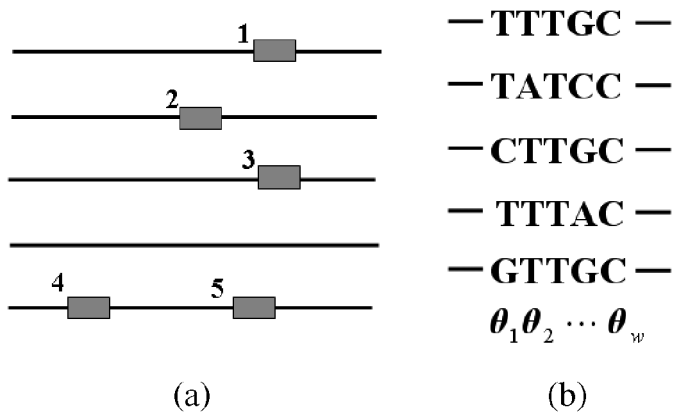

FIG. 8. (a) *The sequences are viewed as a mixture of motif sites* (*the rectangles*) *and background letters*. (b) *Motif model is represented by PWM, or equivalently, a product multinomial distribution, where we assume that the columns within a motif are independent.*

(each $\boldsymbol{\beta}_i$ is a length-4 vector corresponding to A, C, G, T), and a Beta(a, b) prior $\pi(p_0)$ on p_0. In (17), $\mathbf{c}_i$ $(i = 1, 2, \ldots, w)$ is the count vector for the ith position of the motif sites $\mathbf{S}(\mathbf{A})$ [e.g., $\mathbf{c}_1 = (c_{1A}, c_{1C}, c_{1G}, c_{1T})$ counts the total number of A, C, G, T in the first position of the motif in all the sequences], the notation $\boldsymbol{\theta}_i^{(\mathbf{c}_i+\boldsymbol{\beta}_i)} = \prod_j \theta_{ij}^{(c_{ij}+\beta_{ij})}$ $(j = A, C, G, T)$, and $L = \sum_{i=1}^{M} L_i$. Since the main goal is to find the motif binding sites given the sequence, $P(\mathbf{A}|\mathbf{S})$ is of primary interest. The parameters ($\boldsymbol{\Theta}$ and p_0) can thus be integrated out resulting in the following "collapsed" posterior distribution [14, 24]:

$$P(\mathbf{A}|\mathbf{S}) \propto \frac{1}{P(\mathbf{S}(\mathbf{A})|\mathbf{A}, \boldsymbol{\theta}_0)} \times \frac{\Gamma(|\mathbf{A}| + a)\Gamma(L - |\mathbf{A}| + b)}{\Gamma(L + a + b)} \times \prod_{i=1}^{w} \frac{\Gamma(\mathbf{c}_i + \boldsymbol{\beta}_i)}{\Gamma(|\mathbf{A}| + |\boldsymbol{\beta}_i|)}, \tag{18}$$

where $\Gamma(\mathbf{c}_i + \boldsymbol{\beta}_i) = \prod_j \Gamma(c_{ij} + \beta_{ij})$ $(j = A, C, G, T)$ and $|\boldsymbol{\beta}_i| = \sum_j \beta_{ij}$.

6.2.2. *The equi-energy motif sampler.* Our target is to sample from $P(\mathbf{A}|\mathbf{S})$. The EE sampler starts from a set of modified distributions,

$$f_i(\mathbf{A}) \propto \exp\left(-\frac{h(\mathbf{A}) \vee H_i}{T_i}\right), \qquad i = 0, 1, \ldots, K,$$

where $h(\mathbf{A}) = -\log P(\mathbf{A}|\mathbf{S})$ is the energy function for this problem. At the target chain ($i = 0$, $T_0 = 1$), we simply implement the Gibbs sampler, that is, sequentially sample the indicator array $\mathbf{A}$ one position at a time with the rest of the positions fixed. To update other sampling chains at $i = 1, 2, \ldots, K$, given the current sample $\mathbf{A}$ we first estimate the motif pattern $\hat{\boldsymbol{\Theta}}$ by a simple frequency counting with some extra pseudocounts, where the number of pseudocounts increases linearly with the chain index so that at higher-order chains $\hat{\boldsymbol{\Theta}}$ is stretched more toward a uniform weight matrix. Next, we fix $\hat{p}_0 = 1/\bar{L}$, where $\bar{L}$ is the average sequence length. Given $\hat{\boldsymbol{\Theta}}$ and $\hat{p}_0$, we sample each position in the sequences independently to obtain a new indicator array $\mathbf{A}^*$ according to the Bayes rule: Suppose the nucleotides at positions j to $j + w - 1$ in sequence k are $x_1 x_2 \cdots x_w$. We propose $\mathbf{A}^*_{kj} = 1$ with probability

$$q_{kj} = \frac{\hat{p}_0 \prod_{n=1}^{w} \hat{\boldsymbol{\Theta}}_{nx_n}}{\hat{p}_0 \prod_{n=1}^{w} \hat{\boldsymbol{\Theta}}_{nx_n} + (1 - \hat{p}_0) P(x_1 \cdots x_w|\boldsymbol{\theta}_0)},$$

where $P(x_1 \cdots x_w|\boldsymbol{\theta}_0)$ denotes the probability of generating these nucleotides from the background model. Then $\mathbf{A}^*$ is accepted to be the new indicator array according to the Metropolis–Hastings ratio

$$r = \frac{f_i(\mathbf{A}^*)}{f_i(\mathbf{A})} \cdot \frac{P(\mathbf{A}|\hat{\boldsymbol{\Theta}}^*)}{P(\mathbf{A}^*|\hat{\boldsymbol{\Theta}})}, \tag{19}$$

where $P(\mathbf{A}^*|\hat{\boldsymbol{\Theta}}) = \prod_{k,j} q_{kj}$ denotes the proposal probability of generating the sample $\mathbf{A}^*$ given current $\mathbf{A}$. In addition to the above updating, the EE motif sampler performs the equi-energy jump in each iteration with probability p_{ee} to help the sampler move freely between different local modes.

6.2.3. *Sampling motifs in "low complexity" sequences.* In the genomes of higher organisms, the presence of long stretches of simple repeats, such as $AAAA\ldots$ or $CGCGCG\ldots$ often makes the motif discovery more difficult, because these repeated patterns are local traps for the algorithms—even the most popular motif finding algorithms based on the Gibbs sampler, such as BioProspector [26] and AlignACE [33], are often trapped to the repeats and miss the true motif pattern. To test whether the EE motif sampler is capable of finding the motifs surrounded by simple repeats, we constructed a set of sequences with the following transition matrix for the background model:

$$\boldsymbol{\theta}_0 = \begin{bmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{bmatrix}, \tag{20}$$

where $\alpha = 0.12$. The data set contained ten sequences, each of length 200 base pairs (i.e., $L_1 = L_2 = \cdots = L_{10} = 200$). Then we independently generated 20 motif sites from the weight matrix whose logo plot [34] is shown in Figure 9(a) and inserted them into the sequences randomly.

The EE motif sampler was applied to this data set with $w = 12$. To set up the energy and temperature ladder, we randomly picked 15 segments of length w in the sequences, treated them as motif sites, and used the corresponding energy value as the upper bound of the energy H_K. For the lower bound H_0, a rough value was estimated by calculating the energy function (18) for typical motif strength with a reasonable number of true sites. Note that since the EE sampler can adaptively adjust the energy and temperature ladder (see Section 3.3), the bound H_0 does not need to be very precise. After some trials H_0 was set to be -50. We utilized $K + 1 = 5$ chains. The energy ladder H_j was set by a geometric progression between H_0 and H_K. The temperatures were set by $(H_{j+1} - H_j)/T_j = 5$. The equi-energy jump probability p_{ee} was taken to be 0.1. We ran each chain for 1000 iterations. Our algorithm predicted 18.4 true sites (out of 20) with 1.0 false site on average over ten independent simulations—the EE sampler successfully found the true motif pattern to a large extent. The sequence logo plot of the predicted sites is shown in Figure 9(b), which is very close to the true pattern in Figure 9(a) that generates the motif sites.

The performance of the EE motif sampler was compared with that of BioProspector and AlignACE, where we set the maximum number of motifs to be detected to 3 and each algorithm was repeated ten times for the data set. The motif width was set to be $w = 12$. Both algorithms, however, tend to be trapped in

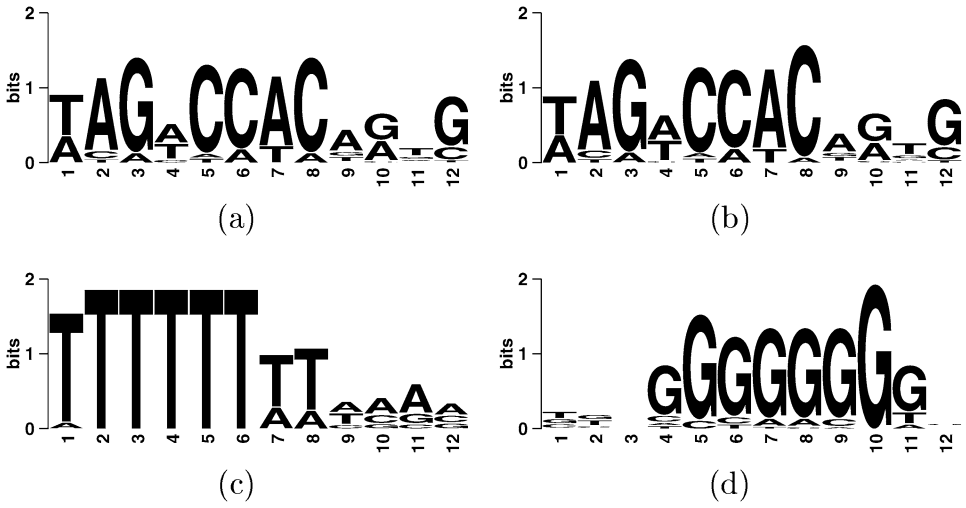

FIG. 9. *Sequence logo plots*. (a) *The motif pattern that generates the simulated sites*. (b) *The pattern of the predicted sites by the EE motif sampler.* (c) *and* (d) *Repetitive patterns found by BioProspector and AlignACE.*

some local modes and output repeats like those shown in Figures 9(c) and (d). One sees from this simple example that the EE sampler, capable of escaping from the numerous local modes, could increase the global searching ability of the sampling algorithm.

6.2.4. *Sampling mixture motif patterns.* If we suspect there are multiple distinct motif patterns in the same set of sequences, one strategy is to introduce more motif matrices, one for each motif type [25]. Alternatively, if we view the different motif patterns as distinct local modes in the sample space, our task is then to design a motif sampler that frequently switches between different modes. This task is almost impossible for the Gibbs sampler, since it can easily get stuck in one local mode (one motif pattern) and have no chance to jump to other patterns in practice. We thus test if the EE sampler can achieve this goal.

In our simulation, we generated 20 sites from each of two different motif models with logo plots in Figures 10(a) and (b) and inserted them randomly into 20 generated background sequences, each of length 100. We thus have 40 motif sites in 20 sequences. The energy ladder in the EE sampler was set by a geometric progression in the range [0, 110] (this is obtained similarly to the previous example). The equi-energy jump probability p_{ee} was set to 0.1. The EE motif sampler used 10 chains; each had 1000 iterations. We recorded the frequency of each position in the sequences being the start of a motif site. Figure 10(c) shows the frequency for the starting positions of the 40 true motif sites (i.e., the probability that each individual motif site was visited by the EE motif sampler). It can be seen that the EE sampler visited both motif patterns with a ratio of 0.4:0.6; by contrast, the Gibbs

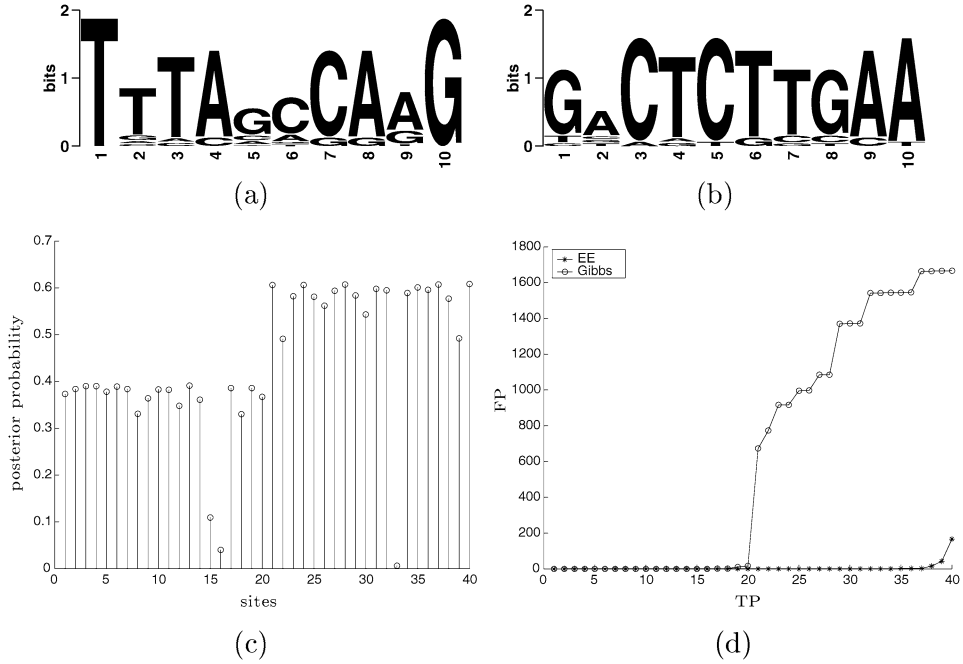

FIG. 10. (a) *and* (b) *Two different motifs inserted in the sequences*. (c) *The marginal posterior distribution of the* 40 *true sites*, *which reports the frequencies that these true site-starting positions are correctly sampled during the iterations*. *Sites* 1–20 *are of one motif type*; *sites* 21–40 *are of the other motif type*. (d) *ROC curve comparison between the EE sampler and the Gibbs sampler.*

motif sampler can only visit one pattern in a single run. We further sorted all the positions according to their frequencies of being a motif-site-start in descending order. For any q between 0 and 1, one could accept any position with frequency greater than q as a predicted motif site of the sampler. Thus by decreasing q from 1 to 0, the numbers of both true and false discovered sites increase, which gives the so-called ROC curve (receiver operating characteristic curve) that plots the number of false positive sites versus the number of true positive sites as q varies. Figure 10(d) shows the ROC curves for both the EE motif sampler and the Gibbs motif sampler.

It can be seen from the figure that for the first 20 true sites, the two samplers showed roughly the same performance. But since the Gibbs sampler missed one mode, the false positive error rate increased dramatically when we further decreased q to include more true sites. By being able to visit both modes (motif patterns), the EE motif sampler, on the other hand, had a very small number of false positive sites until we included 38 true sites, which illustrates that the EE motif sampler successfully distinguished both types of motif patterns from the background sequences.

6.3. *The HP model for protein folding.* Proteins are heteropolymers of 20 types of amino acids. For example, the primary sequence of a protein, cys-ile-leu-lys-glu-met-ser-ile-arg-lys, tells us that this protein is a chain of 10 amino acids linked by strong chemical bonds (peptide bonds) in the backbone. Each different amino acid has a distinct side group that confers different chemical properties to it. Under a normal cellular environment, the side groups form weak chemical bonds (mainly hydrogen bonds) with each other and with the backbone so that the protein can fold into a complex, three-dimensional conformation. Classic experiments such as the ribonuclease refolding experiments [35] suggest that the three-dimensional conformations of most proteins are determined by their primary sequences. The problem of computing the 3D conformation from a primary sequence, known as the protein folding problem, has been a major challenge for biophysicists for over 30 years. This problem is difficult for two reasons. First, there are uncertainties on how to capture accurately all the important physical interactions such as covalent bonds, electrostatic interactions, and interaction between the protein and its surrounding water molecules, and so on, into a single energy function that can be used in minimization and simulation computations. Furthermore, even if the energy function is not in question, we still do not have efficient algorithms for computing the protein conformation from the energy function. For these reasons, biophysicists have developed simplified protein folding models with greatly reduced complexity in both the conformational space and the energy function. It is hoped that the reduced complexity in these simplified models will allow us to deduce insights about the statistical mechanics of the protein folding process through extensive numerical computations.

The HP model is a simplified model that is of great current interest. In this model, there are only two types of amino acids, namely a hydrophilic type (P-type) that is capable of favorable interaction with water molecules, and a hydrophobic type (H-type) that does not interact well with water. Thus, the primary sequence of the length-10 protein in the beginning of this section is simplified to the sequence H-H-H-P-P-H-P-H-P-P. The conformation of the protein chain is specified once the spatial position of each of its amino acids is known. In the HP model space is modeled as a regular lattice in two or three dimensions. Thus the conformation of a length-k protein is a vector $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ where each x_i is a lattice point. Of course, neighbors in the backbone must also be neighbors on the lattice. Figure 11 gives one possible conformation of the above length-10 protein in a two-dimensional lattice model. Just like oil molecules in water, the hydrophobic nature of the H-type amino acids will drive clusters together so as to minimize their exposure to water. Thus we give a favorable energy [i.e., an energy of $\mathcal{E}(x_i, x_j) = -1$] for each pair of H-type amino acids that are not neighbors in the backbone but are lattice neighbors of each other in the conformation [see Figure 11(b)]. All other neighbor pairs are neutral [i.e., having an energy contribution $\mathcal{E}(x_i, x_j) = 0$;

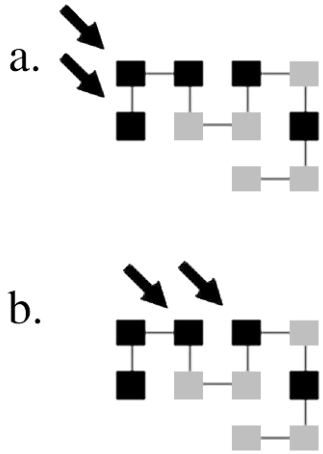

FIG. 11. *One possible conformation of the length-*10 *proteins in a* 2*D lattice. H-type and P-type amino acids are represented by black and gray squares, respectively.*

see Figure 11(a)]. The energy of the conformation is given by

$$h(\mathbf{x}) = \sum_{|i-j|>1} \mathcal{E}(x_i, x_j).$$

Since its introduction by Lau and Dill [19], the HP model has been extensively studied and found to provide valuable insights [4]. We have applied equi-energy sampling to study the HP model in two dimensions. Here we present some summary results to illustrate how our method can provide information that is not accessible to standard MC methods, such as the relative contribution of the entropy to the conformational distribution, and the phase transition from disorder to order. The detailed implementation of the algorithm, the full description and the physical interpretations of the results are available in a separate paper [16].

Table 5 presents estimated (normalized) density of states at various energy levels for a protein of length 20: H-P-H-P-P-H-H-P-H-P-P-H-P-H-H-P-P-H-P-H. For this protein there are 83,770,155 possible conformations with energies ranging from 0 (completed unfolded chain) to -9 (folded conformations having the maximum number of nine H-to-H interactions). The estimated and exact relative frequencies of the energies in this collection of conformations are given in the table. The estimates are based on five independent runs each consisting of one million steps where the probability of proposing an equi-energy move is set to be $p_{\text{ee}} = 0.1$. It is clear that the method yielded very accurate estimates of the density of states at each energy level even though we have sampled only a small proportion of the population of conformations. The equi-energy move is important for the success of the method: if we eliminate these moves, then the algorithm performs very poorly in estimating the entropy at the low-energy end; for example, the estimated density of state at $E = -9$ becomes $(1.545 \pm 4.539) \times 10^{-12}$, which is four orders of magnitude away from the exact value.

TABLE 5
The normalized density of states estimated from the EE sampler compared with the actual value

Energy	Estimated density of states	Actual value	t-value
−9	$(4.751 \pm 2.087) \times 10^{-8}$	4.774×10^{-8}	−0.011
−8	$(1.155 \pm 0.203) \times 10^{-6}$	1.146×10^{-6}	0.043
−7	$(1.452 \pm 0.185) \times 10^{-5}$	1.425×10^{-5}	0.144
−6	$(1.304 \pm 0.189) \times 10^{-4}$	1.237×10^{-4}	0.354
−5	$(9.602 \pm 1.332) \times 10^{-4}$	9.200×10^{-4}	0.302
−4	$(6.365 \pm 0.627) \times 10^{-3}$	6.183×10^{-3}	0.291
−3	$(3.600 \pm 0.228) \times 10^{-2}$	3.514×10^{-2}	0.377
−2	$(1.512 \pm 0.054) \times 10^{-1}$	1.489×10^{-1}	0.423
−1	$(3.758 \pm 0.044) \times 10^{-1}$	3.779×10^{-1}	−0.474
0	$(4.296 \pm 0.071) \times 10^{-1}$	4.309×10^{-1}	−0.181

The t-value is defined as the difference between the estimate and the actual value divided by the standard deviation.

We also use the equi-energy sampler to study phase transition from a disordered state, where the conformational distribution is dominated by the entropy term and the protein is likely to be in an unfolded state with high energy, to an ordered state, where the conformation is likely to be compactly folded structures with low energy. We use the "minimum box size" (BOXSIZE) as a parameter to measure the extent the protein has folded. BOXSIZE is defined as the size of the smallest possible rectangular region containing all the amino acid positions in the conformation. For the 20-length protein, Figure 12 gives a plot of estimated Boltzmann averages of BOXSIZE at ten different temperatures, which are available from a single run of the equi-energy sampler using five energy ranges. We see that there is a rather sharp transition from order (folded state) to disorder at the temperature range $T = 0.25$ to $T = 1$, with an inversion point around $T = 0.5$. In our energy scale, room temperature corresponds to $T = 0.4$ [16]. Thus at room temperature this length-20 protein will not always assume the minimum energy conformation; rather, it still has a significant probability of being in high-energy, unfolded states. It is clear from this example that the equi-energy sampler is capable of providing estimates of many parameters that are important for the understanding of the protein folding problem from a statistical mechanics and thermodynamics perspective.

7. Discussion. We have presented a new Monte Carlo method that is capable of sampling from multiple energy ranges. By using energy truncation and matching temperature and step sizes to the energy range, the algorithm can explore the energy landscape with great flexibility. Most importantly, the algorithm relies on a new type of move—the equi-energy jumps—to reach regions of the sample space that have energy similar to the current state but may be separated by steep energy

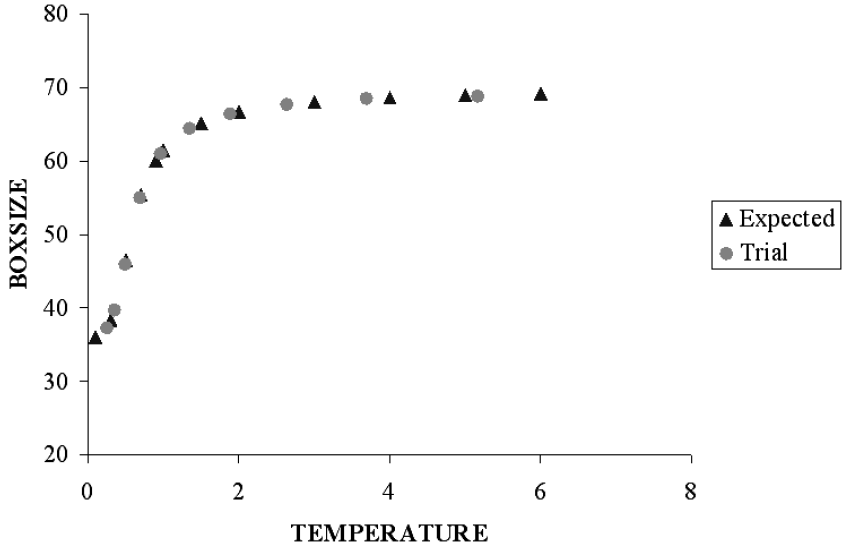

FIG. 12. *The Boltzmann average of BOXSIZE at different temperatures.*

barriers. The equi-energy jumps provide direct access to these regions that are not reachable with classic Metropolis moves.

We have also explained how to obtain estimates of the density of states and the microcanonical averages from the samples generated by the equi-energy sampler. Because of the duality between temperature-dependent parameters and energy-dependent parameters, the equi-energy sampler thus also provides estimates of expectations under any fixed temperature.

Our method has connections with two of the most powerful Monte Carlo methods currently in use, namely parallel tempering and multicanonical sampling. In our numerical examples, equi-energy sampling is seen to be more efficient than parallel tempering and it provides estimates of quantities (e.g., entropy) not accessible by parallel tempering. In this paper we have not provided direct numerical comparison of our method with multicanonical sampling. One potential advantage over multicanonical sampling is that the equi-energy jumps allow us to make use of conformations generated in the higher energy ranges to make movement across energy barriers efficiently. Thus equi-energy sampling "remembers" energy-favorable configurations and makes use of them in later sampling steps. By contrast, multicanonical sampling makes use of previous configurations only in terms of the estimated density of states function. In this sense our method has better memory than multicanonical sampling. It is our belief that equi-energy sampling holds promise for improved sampling efficiency as an all-purpose Monte Carlo method, but definitive comparison with both parallel tempering and multicanonical sampling must await future studies.

In addition to its obvious use in statistical physics, our method will also be useful to statistical inference. It offers an effective means to compute posterior expectations and marginal distributions. Furthermore, the method provides direct estimates of conditional expectations given fixed energy levels (the microcanonical averages). Important information on the nature of the likelihood or the posterior distribution, such as multimodality, can be extracted by careful analysis of these conditional averages. For instance, in Section 4 we see that the equi-energy sampler running on (12) reveals that there is a change point in the density of states as well as the microcanonical averages [see Figures 6(c) and (d)], which clearly indicates the multimodality of the underlying distribution. We believe that the design of methods for inferring the properties of the likelihood surface or the posterior density, based on the output of the equi-energy sampler, will be a fruitful topic for future investigations.

Acknowledgment. The authors thank Jason Oh for computational assistance and helpful discussion. The authors are grateful to the Editor, the Associate Editor and two referees for thoughtful and constructive comments.

REFERENCES

[1] BAILEY, T. L. and ELKAN, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Second International Conference on Intelligent Systems for Molecular Biology* **2** 28–36. AAAI Press, Menlo Park, CA.

[2] BERG, B. A. and NEUHAUS, T. (1991). Multicanonical algorithms for first order phase-transitions. *Phys. Lett. B* **267** 249–253.

[3] BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B* **55** 25–37. MR1210422

[4] DILL, K. A. and CHAN, H. S. (1997). From Levinthal to pathways to funnels. *Nature Structural Biology* **4** 10–19.

[5] EDWARDS, R. G. and SOKAL, A. D. (1988). Generalization of the Fortuin–Kasteleyn–Swendsen–Wang representation and Monte Carlo algorithm. *Phys. Rev. D* (*3*) **38** 2009–2012. MR0965465

[6] GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

[7] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6** 721–741.

[8] GEYER, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics*: *Proc. 23rd Symposium on the Interface* (E. M. Keramidas, ed.) 156–163. Interface Foundation, Fairfax Station, VA.

[9] GEYER, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report 568, School of Statistics, Univ. Minnesota.

[10] GEYER, C. J. and THOMPSON, E. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.

[11] HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

[12] HIGDON, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.* **93** 585–595.

[13] HUKUSHIMA, K. and NEMOTO, K. (1996). Exchange Monte Carlo and application to spin glass simulations. *J. Phys. Soc. Japan* **65** 1604–1608.
[14] JENSEN, S. T., LIU, X. S., ZHOU, Q. and LIU, J. S. (2004). Computational discovery of gene regulatory binding motifs: A Bayesian perspective. *Statist. Sci.* **19** 188–204. MR2082154
[15] KONG, A., LIU, J. S. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.
[16] KOU, S. C., OH, J. and WONG, W. H. (2006). A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling. *J. Chemical Physics* **124** 244903.
[17] KOU, S. C., XIE, X. S. and LIU, J. S. (2005). Bayesian analysis of single-molecule experimental data (with discussion). *Appl. Statist.* **54** 469–506. MR2137252
[18] LANDAU, D. P. and BINDER, K. (2000). *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge Univ. Press. MR1781083
[19] LAU, K. F. and DILL, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22** 3986–3997.
[20] LAWRENCE, C. E., ALTSCHUL, S. F., BOGUSKI, M. S., LIU, J. S., NEUWALD, A. F. and WOOTTON, J. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262** 208–214.
[21] LAWRENCE, C. E. and REILLY, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7** 41–51.
[22] LI, K.-H. (1988). Imputation using Markov chains. *J. Statist. Comput. Simulation* **30** 57–79. MR1005883
[23] LIANG, F. and WONG, W. H. (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Statist. Assoc.* **96** 653–666. MR1946432
[24] LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966. MR1294740
[25] LIU, J. S., NEUWALD, A. F. and LAWRENCE, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc.* **90** 1156–1170.
[26] LIU, X., BRUTLAG, D. L. and LIU, J. S. (2001). BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pacific Symp. Biocomputing* **6** 127–138.
[27] MARINARI, E. and PARISI, G. (1992). Simulated tempering: A new Monte Carlo scheme, *Europhys. Lett.* **19** 451–458.
[28] MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406
[29] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics* **21** 1087–1091.
[30] MIRA, A., MOLLER, J. and ROBERTS, G. (2001). Perfect slice samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 593–606. MR1858405
[31] NEAL, R. M. (2003). Slice sampling (with discussion). *Ann. Statist.* **31** 705–767. MR1994729
[32] ROBERTS, G. and ROSENTHAL, J. S. (1999). Convergence of slice sampler Markov chains. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 643–660. MR1707866
[33] ROTH, F. P., HUGHES, J. D., ESTEP, P. W. and CHURCH, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* **16** 939–945.
[34] SCHNEIDER, T. D. and STEPHENS, R. M. (1990). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research* **18** 6097–6100.

[35] SELA, M., WHITE, F. H. and ANFINSEN, C. B. (1957). Reductive cleavage of disulfide bridges in ribonuclease. *Science* **125** 691–692.
[36] STORMO, G. D. and HARTZELL, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* **86** 1183–1187.
[37] SWENDSEN, R. H. and WANG, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58** 86–88.
[38] TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357
[39] TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762. MR1329166
[40] WANG, F. and LANDAU, D. P. (2001). Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E* **64** 056101.
[41] ZHOU, Q. and WONG, W. H. (2004). CisModule: *De novo* discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA* **101** 12114–12119.

S. C. KOU
Q. ZHOU
DEPARTMENT OF STATISTICS
SCIENCE CENTER
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: kou@stat.harvard.edu
zhou@stat.harvard.edu

W. H. WONG
DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: whwong@stanford.edu

The Annals of Statistics
2006, Vol. 34, No. 4, 1620–1628
DOI: 10.1214/009053606000000489
Main article DOI: 10.1214/009053606000000515

DISCUSSION OF "EQUI-ENERGY SAMPLER" BY KOU, ZHOU AND WONG

BY YVES F. ATCHADÉ AND JUN S. LIU

University of Ottawa and Harvard University

We congratulate Samuel Kou, Qing Zhou and Wing Wong (referred to subsequently as KZW) for this beautifully written paper, which opens a new direction in Monte Carlo computation. This discussion has two parts. First, we describe a very closely related method, multicanonical sampling (MCS), and report a simulation example that compares the equi-energy (EE) sampler with MCS. Overall, we found the two algorithms to be of comparable efficiency for the simulation problem considered. In the second part, we develop some additional convergence results for the EE sampler.

1. A multicanonical sampling algorithm. Here, we take on KZW's discussion about the comparison of the EE sampler and MCS. We compare the EE sampler with a general state-space extension of MCS proposed by Atchadé and Liu [1]. We compare the two algorithms on the multimodal distribution discussed by KZW in Section 3.4.

Let $(\mathcal{X}, \mathcal{B}, \lambda)$ be the state space equipped with its σ-algebra and appropriate measure, and let $\pi(x) \propto e^{-h(x)}$ be the density of interest. Following the notation of KZW, we let $H_0 < H_1 < \cdots < H_{K_e} < H_{K_e+1} = \infty$ be a sequence of energy levels and let $D_j = \{x \in \mathcal{X} : h(x) \in [H_j, H_{j+1})\}$, $0 \le j \le K_e$, be the energy rings. For $x \in \mathcal{X}$, define $I(x) = j$ if $x \in D_j$. Let $1 = T_0 < T_1 < \cdots < T_{K_t}$ be a sequence of "temperatures." We use the notation $k^{(i)}(x) = e^{-h(x)/T_i}$, so that $\pi^{(i)}(x) = k^{(i)}(x)/\int k^{(i)}(x)\lambda(dx)$. Clearly, $\pi^{(0)} = \pi$. We find it more convenient to use the notation $\pi^{(i)}$ instead of π_i as in KZW. Also note that we did not flatten $\pi^{(i)}$ as KZW did.

The goal of our MCS method is to generate a Markov chain on the space $\mathcal{X} \times \{0, 1, \ldots, K_t\}$ with invariant distribution

$$\pi(x, i) \propto \sum_{j=0}^{K_e} \frac{k^{(i)}(x)}{Z_{i,j}} \mathbf{1}_{D_j}(x)\lambda(dx),$$

where $Z_{i,j} = \int k^{(i)}(x)\mathbf{1}_{D_j}(x)\lambda(dx)$. With a well-chosen temperature sequence (T_i) and energy levels (H_j), such a Markov chain would move very easily from any temperature level $\mathcal{X} \times \{i\}$ to another. And inside each temperature level $\mathcal{X} \times \{i\}$, the algorithm would move very easily from any energy ring D_j to another. Unfortunately, the normalizing constants $Z_{i,j}$ are not known. They are estimated as

Received October 2005.

part of the algorithm using the Wang–Landau recursion that we describe below. To give the details, we need a proposal kernel $Q_i(x, dy) = q_i(x, dy)\lambda(dy)$ on $\mathcal{X}$, a proposal kernel $\Delta(i, j)$ on $\{0, \ldots, K_t\}$ and (γ_n), a sequence of positive numbers. We discuss the choice of these parameters later.

ALGORITHM 1.1 (*Multicanonical sampling*).

Initialization. Start the algorithm with some arbitrary $(X_0, t_0) \in \mathcal{X} \times \{0, 1, \ldots, K_t\}$. For $i = 0, \ldots, K_t$, $j = 0, \ldots, K_e$ we set all the weights to $\phi_0^{(i)}(j) = 1$.

Recursion. Given $(X_n, t_n) = (x, i)$ and $(\phi_n^{(i)}(j))$, flip a θ-coin.

If Tail. Sample $Y \sim Q_i(x, \cdot)$. Set $X_{n+1} = Y$ with probability $\alpha^{(i)}(x, Y)$; otherwise set $X_{n+1} = x$, where

$$\alpha^{(i)}(x, y) = \min\left[1, \frac{k^{(i)}(y)}{k^{(i)}(x)} \frac{\phi_n^{(i)}(I(x))}{\phi_n^{(i)}(I(y))} \frac{q_i(y, x)}{q_i(x, y)}\right]. \tag{1.1}$$

Set $t_{n+1} = i$.

If Head. Sample $j \sim \Delta(i, \cdot)$. Set $t_{n+1} = j$ with probability $\beta_x(i, j)$; otherwise set $t_{n+1} = i$, where

$$\beta_x(i, j) = \min\left[1, \frac{k^{(j)}(x)}{k^{(i)}(x)} \frac{\phi_n^{(i)}(I(x))}{\phi_n^{(j)}(I(x))} \frac{\Delta(j, i)}{\Delta(i, j)}\right]. \tag{1.2}$$

Set $X_{n+1} = x$.

Update the weights. Write $(t_{n+1}, I(X_{n+1})) = (i_0, j_0)$. Set

$$\phi_{n+1}^{(i_0)}(j_0) = \phi_n^{(i_0)}(j_0)(1 + \gamma_n), \tag{1.3}$$

leaving the other weights unchanged.

If we choose $K_t = 0$ in the algorithm above we obtain the MCS of [9] (the first MCS algorithm is due to [4]) and $K_e = 0$ gives the simulated tempering algorithm of [6]. The recursion (1.3) is where the weights $Z_{i,j}$ are being estimated. Note that the resulting algorithm is no longer Markovian. Under some general assumptions, it is shown in [1] that $\theta_n^{(i)}(j) := \frac{\phi_n^{(i)}(j)}{\sum_{i=0}^{K_t} \sum_{l=0}^{K_e} \phi_n^{(i)}(l)} \to \int_{D_j} \pi^{(i)}(x)\lambda(dx)$ as $n \to \infty$.

The MCS can be seen as a random-scan-Gibbs sampler on the two variables $(x, i) \in \mathcal{X} \times \{0, \ldots, K_t\}$, so the choice $\theta = 1/2$ for coin flipping works well. The proposal kernels Q_i can be chosen as in a standard Metropolis–Hastings algorithm. But one should allow Q_i to make larger proposal moves for larger i (i.e., hotter distributions). The proposal kernel Δ can be chosen as a random

walk on $\{0, \ldots, K_t\}$ (with reflection on 0 and K_t). In our simulations, we use $\Delta(0,1) = \Delta(K_t, K_t - 1) = 1$, $\Delta(i, i-1) = \Delta(i, i+1) = 1/2$ for $i \notin \{0, K_t\}$.

It can be shown that the sequence (θ_n) defined above follows a stochastic approximation with step size (γ_n). So choosing (γ_n) is the same problem as choosing a step-size sequence in a stochastic approximation algorithm. We follow the new method proposed by Wang and Landau [9] where (γ_n) is selected adaptively. Wang and Landau's idea is to monitor the convergence of the algorithm and adapt the step size accordingly. We start with some initial value γ_0 and (γ_n) is defined by $\gamma_n = (1+\gamma_0)^{1/(k+1)} - 1$ for $\tau_k < n \le \tau_{k+1}$, where $0 = \tau_0 < \tau_1 < \cdots$ is a sequence of stopping times. Assuming τ_i finite, τ_{i+1} is the next time $k > \tau_i$ where the occupation measures (obtained from time $\tau_i + 1$ on) of all the energy rings in all the temperature levels are approximately equal. Various rules can be used to check that the occupation measures are approximately equal. Following [9], we check that the smallest occupation measure obtained is greater than c times the mean occupation, where c is some constant (e.g., $c = 0.2$) that depends on the complexity of the sampling problem.

It is an interesting question to know whether this method of choosing the step-size sequence can be extended to more general stochastic approximation algorithms. A theoretical justification of the efficiency of the method is also an open question.

2. Comparison of EE sampler and MCS. To use MCS to estimate integrals of interest such as $\mathbb{E}_{\pi_0}(g(X))$, one can proceed as KZW did by writing $\mathbb{E}_{\pi_0}(g(X)) = \sum_{j=0}^{K_e} p_j \mathbb{E}_{\pi_0}(g(X)|X \in D_j)$. Samples from the high-temperature chains can be used to estimate the integrals $\mathbb{E}_{\pi_0}(g(X)|X \in D_j)$ by importance reweighting in the same way as KZW did. In the case of MCS, the probabilities $p_j = \Pr_{\pi_0}(X \in D_j)$ are estimated by $\hat{p}_j = \frac{\phi_n^{(0)}(j)}{\sum_{l=0}^{K_e} \phi_n^{(0)}(l)}$.

We compared the performances of the EE sampler and the MCS described above for the multimodal example in Section 3.4 of KZW. To make the two samplers comparable, each chain in the EE sampler was run for N iterations. We did the simulations for $N = 10^4$, $N = 5 \times 10^4$ and $N = 10 \times 10^4$. For the MC sampler, we used $K_t = K_e = K$ and the algorithm was run for $(K+1) \times N$ total iterations. We repeated each sampler for $n = 100$ iterations in order to estimate the finite sample standard deviations of the estimates they provided. Table 1 gives the improvements (in percentage) of MCS over EE sampling. $\Pr_\pi(X \in B)$ is the probability under π of the union of all the discs with centers μ_i (the means of the mixture) and radius $\sigma/2$. As we can see, when estimating global functions such as moments of the distribution, the two samplers have about the same accuracy with a slight advantage for MCS. But the EE sampler outperformed MCS when estimating $\Pr_\pi(X \in B)$. The MCS is an importance sampling algorithm with a stationary distribution that is more widespread than π. This may account for the better performance obtained by the EE sampler on $\Pr_\pi(X \in B)$. More thorough empirical and theoretical analyses are apparently required to reach any firmer conclusions.

TABLE 1
Improvement of MCS over EE as given by $(\hat{\sigma}_{\mathrm{EE}}(g) - \hat{\sigma}_{\mathrm{MC}}(g))/\hat{\sigma}_{\mathrm{MC}}(g) \times 100$

	$\mathbb{E}(X_1)$	$\mathbb{E}(X_2)$	$\mathbb{E}(X_2^2)$	$\mathbf{Pr}_{\pi}(X \in B)$
$N = 10^4$	13.77	12.53	8.98	−63.49
$N = 5 \times 10^4$	6.99	−7.31	−10.15	−51.22
$N = 10^5$	1.92	5.79	4.99	−55.31

*The comparisons are based on 100 replications of the samplers for each N.

3. Ergodicity of the equi-energy sampler. In this section we take a more technical look at the EE algorithm and derive some ergodicity results. First, we would like to mention that in the proof of Theorem 2, it is not clear to us how KZW derive the convergence in (5). Equation (5) implicitly uses some form of convergence of the distribution of $X_n^{(i+1)}$ to $\pi^{(i+1)}$ as $n \to \infty$ and it is not clear to us how that follows from the assumption that $\Pr(X_{n+1}^{(i+1)} \in A | X_n^{(i+1)} = x) \to S^{(i+1)}(x, A)$ as $n \to \infty$ for all x, all A.

In the analysis below we fix that problem, but under a more stringent assumption. To state our result, let $(\mathcal{X}, \mathcal{B})$ be the state space of each of the equi-energy chains. If P_1 and P_2 are two transition kernels on $\mathcal{X}$, the product $P_1 P_2$ is also a transition kernel defined as $P_1 P_2(x, A) = \int P_1(x, dy) P_2(y, A)$. Recursively, we define P_1^n as $P_1^1 = P_1$ and $P_1^n = P_1^{n-1} P_1$. If f is a measurable real-valued function on $\mathcal{X}$ and μ is a measure on $\mathcal{X}$, we denote $Pf(x) := \int P(x, dy) f(y)$ and $\mu(f) := \int \mu(dx) f(x)$. Also, for $c \in (0, \infty)$ we write $|f| \le c$ to mean $|f(x)| \le c$ for all $x \in \mathcal{X}$. We define the following distance between P_1 and P_2:

$$|||P_1 - P_2||| := \sup_{x \in \mathcal{X}} \sup_{|f| \le 1} |P_1 f(x) - P_2 f(x)|, \tag{3.1}$$

where the supremum is taken over all $x \in \mathcal{X}$ and over all measurable functions $f : \mathcal{X} \to \mathbb{R}$ with $|f| \le 1$. We say that the transition kernel P is uniformly geometrically ergodic if there exists $\rho \in (0, 1)$ such that

$$|||P^n - \pi||| = O(\rho^n). \tag{3.2}$$

It is well known that (3.2) holds if and only if there exist $\varepsilon > 0$, a nontrivial probability measure ν and an integer $m \ge 1$ such that the so-called $M(m, \varepsilon, \nu)$ minorization condition holds, that is, $P^m(x, A) \ge \varepsilon \nu(A)$ for all $x \in \mathcal{X}$ and $A \in \mathcal{B}$ (see, e.g., [8], Proposition 2). We recall that $T_{\mathrm{MH}}^{(i)}$ denotes the Metropolis–Hastings kernel in the EE sampler. The following result is true for the EE sampler.

THEOREM 3.1. *Assume that* $\forall\, i \in \{0, \ldots, K\}$, $T_{\mathrm{MH}}^{(i)}$ *satisfies a* $M(1, \varepsilon_i, \pi^{(i)})$ *minorization condition and that condition* (iii) *of Theorem* 2 *of the paper holds.*

Then for any bounded measurable function f, as $n \to \infty$,

$$\mathbb{E}\big[f\big(X_n^{(i)}\big)\big] \longrightarrow \pi^{(i)}(f) \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n} f\big(X_k^{(i)}\big) \xrightarrow{a.s.} \pi^{(i)}(f). \tag{3.3}$$

For example, if $\mathcal{X}$ is a compact space and $e^{-h(x)}$ remains bounded away from 0 and ∞, then (3.3) holds. Note that the ith chain in the EE sampler is actually a nonhomogeneous Markov chain with transition kernels $K_0^{(i)}, K_1^{(i)}, \ldots$, where $K_n^{(i)}(x, A) := \Pr[X_{n+1}^{(i)} \in A | X_n^{(i)} = x]$. As pointed out by KZW, for any $x \in \mathcal{X}$ and $A \in \mathcal{B}$, $K_n^{(i)}(x, A) \to S^{(i)}(x, A)$ as $n \to \infty$, where $S^{(i)}$ is the limit transition kernel in the EE sampler. This setup brings to mind the following convergence result for nonhomogeneous Markov chains (see [5], Theorem V.4.5):

THEOREM 3.2. *Let $P, P_0, P_1, \ldots$ be a sequence of transition kernels on $(\mathcal{X}, \mathcal{B})$ such that $\|\!| P_n - P |\!\| \to 0$ and P is uniformly geometrically ergodic with invariant distribution π. Then the Markov chain with transition kernels (P_i) is strongly ergodic; that is, for any initial distribution μ,*

$$\|\!| \mu P_0 P_1 \cdots P_n - \pi |\!\| \to 0 \qquad \text{as } n \to \infty. \tag{3.4}$$

The difficulty in applying this theorem to the EE sampler is that we do not have $\|\!| K_n^{(i)} - S^{(i)} |\!\| \to 0$ but only a setwise convergence $|K_n^{(i)}(x, A) - S^{(i)}(x, A)| \to 0$ for each $x \in \mathcal{X}$, $A \in \mathcal{B}$. The solution we propose is to extend Theorem 3.2 as follows.

THEOREM 3.3. *Let $P, P_0, P_1, \ldots$ be a sequence of transition kernels on $(\mathcal{X}, \mathcal{B})$ such that*:

(i) *For any $x \in \mathcal{X}$ and $A \in \mathcal{B}$, $P_n(x, A) \to P(x, A)$ as $n \to \infty$.*

(ii) *P has invariant distribution π and P_n has invariant distribution π_n. There exists $\rho \in (0, 1)$ such that $\|\!| P^k - \pi |\!\| = O(\rho^k)$ and $\|\!| P_n^k - \pi_n |\!\| = O(\rho^k)$.*

(iii) *$\|\!| P_n - P_{n-1} |\!\| \le O(n^{-\lambda})$ for some $\lambda > 0$.*

Then, if (X_n) is an $\mathcal{X}$-valued Markov chain with initial distribution μ and transition kernels (P_n), for any bounded measurable function f we have

$$\mathbb{E}[f(X_n)] \longrightarrow \pi(f) \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{a.s.} \pi(f) \qquad \text{as } n \to \infty. \tag{3.5}$$

We believe that this result can be extended to the more general class of V-geometrically ergodic transition kernels and then one could weaken the uniform minorization assumption on $T_{\mathrm{MH}}^{(i)}$ in Theorem 3.1. But the proof will certainly be more technical. We now proceed to the proofs of the theorems. We first prove Theorem 3.3 and use it to prove Theorem 3.1.

PROOF OF THEOREM 3.3. It can be easily shown from (ii) that $|||\pi_n - \pi_{n-1}||| \le \frac{1}{1-\rho}|||P_n - P_{n-1}|||$. Therefore, Theorems 3.1 and 3.2 of [2] apply and assert that for any bounded measurable function f, $\mathbb{E}[f(X_n) - \pi_n(f)] \to 0$ and $\frac{1}{n}\sum_{k=1}^{n}[f(X_k) - \pi_k(f)] \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$. To finish, we need to prove that $\pi_n(f) \to \pi(f)$ as $n \to \infty$. To this end, we need the following technical lemma proved in [7], Chapter 11, Proposition 18.

LEMMA 3.1. *Let (f_n) be a sequence of measurable functions and let $\mu, \mu_1, \ldots$ be a sequence of probability measures such that $|f_n| \le 1$ and $f_n \to f$ pointwise $[f_n(x) \to f(x)$ for all $x \in \mathcal{X}]$ and $\mu_n \to \mu$ setwise $[\mu_n(A) \to \mu(A)$ for all $A \in \mathcal{B}]$. Then $\int f_n(x)\mu_n(dx) \to \int f(x)\mu(dx)$.*

Here is how to prove that $\pi_n(f) \to \pi(f)$ as $n \to \infty$. By (i), we have $P_n f(x) \to Pf(x)$ for all $x \in \mathcal{X}$. Then, by (i) and Lemma 3.1, $P_n^2 f(x) = P_n(P_n f)(x) \to P^2 f(x)$ as $n \to \infty$. By recursion, for any $x \in \mathcal{X}$ and $k \ge 1$, $P_n^k f(x) \to P^k f(x)$ as $n \to \infty$. Now, write

$$\begin{aligned} |\pi_n(f) - \pi(f)| &\le |\pi_n(f) - P_n^k f(x)| + |P_n^k f(x) - P^k f(x)| \\ &\quad + |P^k f(x) - \pi(f)| \\ &\le 2\rho^k \sup_{x\in\mathcal{X}} |f(x)| + |P_n^k f(x) - P^k f(x)| \qquad \text{[by (ii)]}. \end{aligned} \tag{3.6}$$

Since $|P_n^k f(x) - P^k f(x)| \to 0$, we see that $|\pi_n(f) - \pi(f)| \to 0$. □

PROOF OF THEOREM 3.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability triplet on which the equi-energy process is defined and let $\mathbb{E}$ be its expectation operator. The result is clearly true by assumption for $i = K$. Assuming that it is true for the $(i+1)$st chain, we will prove it for the ith chain.

The random process $(X_n^{(i)})$ is a nonhomogeneous Markov chain with transition kernel $K_n^{(i)}(x, A) := \Pr[X_{n+1}^{(i)} \in A | X_n^{(i)} = x]$. For any bounded measurable function f, $K_n^{(i)}$ operates on f as follows:

$$K_n^{(i)} f(x) = (1 - p_{\text{ee}}) T_{\text{MH}}^{(i)} f(x) + p_{\text{ee}} \mathbb{E}\big[R_n^{(i)} f(x)\big],$$

where $R_n^{(i)} f(x)$ is a ratio of empirical sums of the $(i+1)$st chain of the form

$$\begin{aligned} R_n^{(i)} f(x) &= \frac{\sum_{k=-N}^{n} \mathbf{1}_{D_{I(x)}}(X_k^{(i+1)}) \alpha^{(i)}(x, X_k^{(i+1)}) f(X_k^{(i+1)})}{\sum_{k=-N}^{n} \mathbf{1}_{D_{I(x)}}(X_k^{(i+1)})} \\ &\quad + f(x) \frac{\sum_{k=-N}^{n} \mathbf{1}_{D_{I(x)}}(X_k^{(i+1)})(1 - \alpha^{(i)}(x, X_k^{(i+1)}))}{\sum_{k=-N}^{n} \mathbf{1}_{D_{I(x)}}(X_k^{(i+1)})} \end{aligned} \tag{3.7}$$

[take $R_n^{(i)} f(x) = 0$ and $p_{\text{ee}} = 0$ when $\sum_{k=-N}^{n} \mathbf{1}_{D_{I(x)}}(X_k^{(i+1)}) = 0$], where $\alpha^{(i)}(x, y)$ is the acceptance probability $\min[1, \frac{\pi^{(i)}(y)\pi^{(i+1)}(x)}{\pi^{(i)}(x)\pi^{(i+1)}(y)}]$. N is how long the $(i+1)$st chain has been running before the ith chain started. Because (3.3) is assumed true for the $(i+1)$st chain and condition (iii) of Theorem 2 of the paper holds, we can assume in the sequel that $\sum_{k=-N}^{n} \mathbf{1}_{D_{I(x)}}(X_k^{(i+1)}) \geq 1$ for all $n \geq 1$. We prove the theorem through a series of lemmas.

LEMMA 3.2. *For the EE sampler, assumption* (i) *of Theorem* 3.3 *holds true.*

PROOF. Because (3.3) is assumed true for the $(i+1)$st chain, the strong law of large numbers and Lebesgue's dominated convergence theorem apply to $R_n^{(i)} f(x)$ and assert that for all $x \in \mathcal{X}$ and $A \in \mathcal{B}$, $K_n^{(i)}(x, A) \to S^{(i)}(x, A)$ as $n \to \infty$, where $S^{(i)}(x, A) = (1 - p_{\text{ee}}) T_{\text{MH}}^{(i)}(x, A) + p_{\text{ee}} \sum_{j=0}^{K} T_{\text{EE}}^{(i,j)}(x, A) \mathbf{1}_{D_j}(x)$, where $T_{\text{EE}}^{(i,j)}$ is the transition kernel of the Metropolis–Hastings with proposal distribution

$$\pi^{(i+1)}(y) \mathbf{1}_{D_j}(y) / p_j^{(i+1)}$$

and invariant distribution

$$\pi^{(i)}(x) \mathbf{1}_{D_j}(x) / p_j^{(i)}. \qquad \square$$

LEMMA 3.3. *For the EE sampler, assumption* (ii) *of Theorem* 3.3 *holds.*

PROOF. Clearly, the minorization condition on $T_{\text{MH}}^{(i)}$ transfers to $K_n^{(i)}$. It then follows that each $K_n^{(i)}$ admits an invariant distribution $\pi_n^{(i)}$ and is uniformly geometrically ergodic toward $\pi_n^{(i)}$ with a rate $\rho_i = 1 - (1 - p_{\text{ee}})\varepsilon_i$. The limit transition kernel $S^{(i)}$ in the EE sampler as detailed above has invariant distribution $\pi^{(i)}$ and also inherits the minorization condition on $T_{\text{MH}}^{(i)}$. $\square$

LEMMA 3.4. *For the EE sampler, assumption* (iii) *of Theorem* 3.3 *holds true with* $\lambda = 1$.

PROOF. Any sequence (x_n) of the form $x_n = \frac{\sum_{k=1}^{n} \alpha_n u_n}{\sum_{k=1}^{n} \alpha_n}$ can always be written recursively as $x_n = x_{n-1} + \frac{\alpha_n}{\sum_{k=1}^{n} \alpha_k}(u_n - x_{n-1})$. Using this, we easily have the bound

$$\left| K_n^{(i)} f(x) - K_{n-1}^{(i)} f(x) \right| \leq 2\mathbb{E}\left[\frac{1}{\sum_{k=-N}^{n} \mathbf{1}_{D_{I(x)}}(X_k^{(i+1)})} \right]$$

for all $x \in \mathcal{X}$ and $|f| \leq 1$. Therefore, the lemma will be proved if we can show that

$$\sup_{0 \leq j \leq K} \mathbb{E}\left[\frac{n}{\sum_{k=-N}^{n} \mathbf{1}_{D_j}(X_k^{(i+1)})} \right] = O(1). \tag{3.8}$$

To do so, we fix $j \in \{0, \ldots, K\}$ and take $\varepsilon \in (0, \delta)$, where $\delta = (1 - p_{\mathrm{ee}})\varepsilon_{i+1} \times \pi^{(i+1)}(D_j) > 0$. We have

$$\begin{aligned}
&\mathbb{E}\left[\frac{n}{\sum_{k=-N}^{n} \mathbf{1}_{D_j}(X_k^{(i+1)})}\right] \\
&\quad = \mathbb{E}\left[\frac{n}{\sum_{k=-N}^{n} \mathbf{1}_{D_j}(X_k^{(i+1)})} \mathbf{1}_{\{\sum_{k=-N}^{n} \mathbf{1}_{D_j}(X_k^{(i+1)}) > n(\delta-\varepsilon)\}}\right] \\
&\qquad + \mathbb{E}\left[\frac{n}{\sum_{k=-N}^{n} \mathbf{1}_{D_j}(X_k^{(i+1)})} \mathbf{1}_{\{\sum_{k=-N}^{n} \mathbf{1}_{D_j}(X_k^{(i+1)}) \le n(\delta-\varepsilon)\}}\right].
\end{aligned} \tag{3.9}$$

The first term on the right-hand side of (3.9) is bounded by $1/(\delta - \varepsilon)$.

The second term is bounded by

$$\begin{aligned}
&n \Pr\left[\sum_{k=-N}^{0} \mathbf{1}_{D_j}(X_k^{(i+1)}) + \sum_{k=1}^{n} \bigl(\mathbf{1}_{D_j}(X_k^{(i+1)}) - \delta\bigr) \le -n\varepsilon\right] \\
&\quad \le n \Pr\bigl[M_n^{(i+1)} \ge n\varepsilon\bigr],
\end{aligned} \tag{3.10}$$

where $M_n^{(i+1)} = \sum_{k=1}^{n} K_{k-1}^{(i+1)}(X_{k-1}^{(i+1)}, D_j) - \mathbf{1}_{D_j}(X_k^{(i+1)})$. For the inequality in (3.10), we use the minorization condition $K_{k-1}^{(i+1)}(x, D_j) \ge \delta$. Now, the sequence $(M_n^{(i+1)})$ is a martingale with increments bounded by 1. By Azuma's inequality ([3], Lemma 1), we have $n \Pr[M_n^{(i+1)} \ge n\varepsilon] \le n \exp(-n\varepsilon^2/2) \to 0$ as $n \to \infty$. □

Theorem 3.1 now follows from Theorem 3.3. □

REFERENCES

[1] ATCHADÉ, Y. F. and LIU, J. S. (2004). The Wang–Landau algorithm for Monte Carlo computation in general state spaces. Technical report.

[2] ATCHADÉ, Y. F. and ROSENTHAL, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11** 815–828. MR2172842

[3] AZUMA, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Math. J.* (2) **19** 357–367. MR0221571

[4] BERG, B. A. and NEUHAUS, T. (1992). Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.* **68** 9–12.

[5] ISAACSON, D. L. and MADSEN, R. W. (1976). *Markov Chains*: *Theory and Applications*. Wiley, New York. MR0407991

[6] MARINARI, E. and PARISI, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* **19** 451–458.

[7] ROYDEN, H. L. (1963). *Real Analysis*. Collier-Macmillan, London. MR0151555

[8] TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762. MR1329166

[9] WANG, F. and LANDAU, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86** 2050–2053.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF OTTAWA
585 KING EDWARD STREET
OTTAWA, ONTARIO
CANADA K1N 6N5
E-MAIL: yatchade@uottawa.ca

DEPARTMENT OF STATISTICS
SCIENCE CENTER
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: jliu@stat.harvard.edu

The Annals of Statistics
2006, Vol. 34, No. 4, 1629–1635
DOI: 10.1214/009053606000000498
Main article DOI: 10.1214/009053606000000515

DISCUSSION OF "EQUI-ENERGY SAMPLER" BY KOU, ZHOU AND WONG

BY MING-HUI CHEN AND SUNGDUK KIM

University of Connecticut

1. Introduction. We first would like to congratulate the authors for their interesting paper on the development of the innovative equi-energy (EE) sampler. The EE sampler provides a solution, which may be better than existing methods, to a challenging MCMC sampling problem, that is, sampling from a multimodal target distribution $\pi(x)$. The EE sampler can be understood as follows. In the equi-energy jump step, (i) points may move within the same mode; or (ii) points may move between two modes; but (iii) points cannot move from one energy ring to another energy ring. In the Metropolis–Hastings (MH) step, points move locally. Although in the MH step, points may not be able to move freely from one mode to another mode, the MH step does help a point to move from one energy ring to another energy ring locally. To maintain certain balance between these two types of operations, an EE jump probability p_{ee} must be specified. Thus, the MH move and the equi-energy jump play distinct roles in the EE sampler. This unique feature makes the EE sampler quite attractive in sampling from a multimodal target distribution.

2. Tuning and "black-box." The performance of the EE sampler depends on the number of energy and temperature levels, K, energy levels $H_0 < H_1 < \cdots < H_K < H_{K+1} = \infty$, temperature ladders $1 = T_0 < T_1 < \cdots < T_k$, the MH proposal distribution, the proposal distribution used in the equi-energy jump step and the equi-energy jump probability p_{ee}. Based on our experience in testing the EE sampler, we felt that the choice of the H_k, the MH proposal and p_{ee} are most crucial for obtaining an efficient EE sampler. In addition, the choice of these parameters is problem-dependent. To achieve fast convergence and good mixing, the EE sampler requires extensive tuning of K, H_k, MH proposal and p_{ee} in particular. A general sampler is designed to be "black box" in the sense that the user need not tune the sampler to the problem. Some attempts have been made for developing such "black-box" samplers in the literature. Neal [4] developed variations on slice sampling that can be used to sample from any continuous distributions and that require little or no tuning. Chen and Schmeiser [2] proposed the random-direction interior-point (RDIP) sampler. RDIP samples from the uniform distribution defined over the region $U = \{(x, y) : 0 < y < \pi(x)\}$ below the curve of the surface defined by $\pi(x)$, which is essentially the same idea used in slice sampling.

Received November 2005.

3. Boundedness. It is not clear why the target distribution $\pi(x)$ must be bounded. Is this a necessary condition required in Theorem 2? It appears that the condition $\sup_x \pi(x) < \infty$ is used only in the construction of energy levels H_k for $k > 0$ for convenience. Would it be possible to relax such an assumption? Otherwise, the EE sampler cannot be applied to sampling from an unbounded $\pi(x)$ such as a gamma distribution with shape parameter less than 1.

If we rewrite

$$D_j = \{x : h(x) \in [H_j, H_{j+1})\} = \big\{x : \pi(x) \in \big(\exp(-H_{j+1}), \exp(-H_j)\big]\big\},$$

we can see that D_0 corresponds to the highest-density region. Thus, if H_1 is appropriately specified, and the guideline given in Section 3.3 is applied to the choice of the rest of the H_j's, the boundedness assumption on $\pi(x)$ may not be necessary.

4. Efficiency. The proposed EE sampler requires $K(B + N)$ iterations before it starts the lowest-order chain $\{X_n^{(0)}, n \geq 0\}$. Note that here B is the number of "burn-in" iterations and N is the number of iterations used in constructing an empirical energy ring $\hat{D}_j^k$. As it is difficult to determine how quickly a Markov chain $\{X_n^{(k)}\}$ converges, a relatively large B may be needed. If the chain $X^{(k)}$ does not converge, the acceptance probability given in Section 3.1 for the equi-energy move at energy levels lower than k may be problematic. Therefore, the EE sampler is quite inefficient as a large number of "burn-in" iterations will be wasted. This may be particularly a problem when K is large. Interestingly, the authors never disclosed what B and N were used in their illustrative examples. Thus, the choice of B and N should be discussed in Section 3.3.

5. Applicability in high-dimensional problems. Based on the guideline of the practical implementation provided in the paper, the number of energy levels K could be roughly proportional to the dimensionality of the target distribution. Thus, for a high-dimensional problem, K could be very large. As a result, the EE sampler may become more inefficient as more "burn-in" iterations are required and at the same time, it may be difficult to tune the parameters involved in the EE sampler.

For example, consider a skewed link model for binary response data proposed by Chen, Dey and Shao [1]. Let $(y_1, y_2, \ldots, y_n)'$ denote an $n \times 1$ vector of n independent dichotomous random variables. Let $x_i = (x_{i1}, \ldots, x_{ip})'$ be a $p \times 1$ vector of covariates. Also let $(w_1, w_2, \ldots, w_n)'$ be a vector of independent latent variables. Then, the skewed link model is formulated as follows: $y_i = 0$ if $w_i < 0$ and 1 if $w_i \geq 0$, where $w_i = x_i'\beta + \delta z_i + \varepsilon_i$, $z_i \sim G$, $\varepsilon_i \sim F$, z_i and ε_i are independent, $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients, δ is the skewness parameter, G is a known cumulative distribution function (c.d.f.) of a skewed distribution, and F is a known c.d.f. of a symmetric distribution. To carry out Bayesian inference for this binary regression model with a skewed link, we

need to sample from the joint posterior distribution of $((w_i, z_i), i = 1, \ldots, n, \beta, \delta)$ given the observed data D. The dimension of the target distribution is $2n + p + 1$. When the sample size n is large, we face a high-dimensional problem. Notice that the dimension of the target distribution can be reduced considerably if we integrate out (w_i, z_i) from the likelihood function. However, in this case, the resulting posterior distribution $\pi(\beta, \delta|D)$ contains many analytically intractable integrals, which could make the EE sampler expensive or even infeasible to implement. The skewed link model is only a simple illustration of a high-dimensional problem. Sampling from the posterior distribution under nonlinear mixed-effects models with missing covariates considered in [5] could be even more challenging.

In contrast, the popular Gibbs sampler may be more attractive and perhaps more suitable for a high-dimensional problem because the Gibbs sampler requires only sampling from low-dimensional conditional distributions. As MH sampling can be embedded into a Gibbs step, would it be possible to develop an EE-within Gibbs sampler?

6. Statistical estimation. In the paper, the authors proposed a sophisticated but interesting Monte Carlo method to estimate the expectation $E_{\pi_0}[g(X)]$ under the target distribution $\pi_0(x) = \pi(x)$ using all chains from the EE sampler. Due to the nature of the EE sampler, the state space $\mathcal{X}$ is partitioned according to the energy levels, that is, $\mathcal{X} = \bigcup_{j=0}^{K} D_j$. Thus, this may be an ideal scenario for applying the partition-weighted Monte Carlo method proposed by Chen and Shao [3]. Let $\{X_i^{(0)}, i = 1, 2, \ldots, n\}$ denote the sample under the chain $X^{(0)}$ $(T = 1)$. Then, the partition-weighted Monte Carlo estimator is given by

$$\hat{E}_{\pi_0}[g(X)] = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{K} w_j g\big(X_i^{(0)}\big) 1\{X_i^{(0)} \in D_j\},$$

where the indicator function $1\{X_i^{(0)} \in D_j\} = 1$ if $X_i^{(0)} \in D_j$ and 0 otherwise, and w_j is the weight assigned to the jth partition. The weights w_j may be estimated using the combined sample, $\{X^{(k)}, k = 1, 2, \ldots, K\}$, under the π_k for $k = 1, 2, \ldots, K$.

7. Example 1. We consider sampling from a two-dimensional normal mixture,

$$f(x) = \sum_{i=1}^{2} \frac{1}{2}\left[\frac{1}{2\pi}|\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i)\right\}\right], \tag{7.1}$$

where

$$x = (x_1, x_2)', \qquad \mu_1' = (0, 0), \qquad \mu_2' = (5, 5)$$

and

$$\Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_i \\ \sigma_1\sigma_2\rho_i & \sigma_2^2 \end{pmatrix}$$

with $\sigma_1 = \sigma_2 = 1.0$, $\rho_1 = 0.99$ and $\rho_2 = -0.99$. The purpose of this example is to examine performance of the EE sampler under a bivariate normal distribution with a high correlation between X_1 and X_2. Since the minimum value of the energy function $h(x) = -\log(f(x))$ is around $\log(4\pi\sigma_1\sigma_2\sqrt{1.0-\rho_i^2}) \approx 0.573$, we took $H_0 = 0.5$. K was set to 2. The energy ladder was set between $H_{\min}$ and $H_{\min} + 100$ in a geometric progression, and the temperatures were between 1 and 60. The equi-energy jump probability p_{ee} was taken to be 0.1. The initial states of the chain $X^{(i)}$ were drawn uniformly from $[0, 1]^2$. The MH proposal was taken to be bivariate Gaussian: $X_{n+1}^{(i)} \sim N_2(X_n^{(i)}, \tau_i^2 T_i I_2)$, where the MH proposal step size τ_i for the ith-order chain $X^{(i)}$ was taken to be 0.5 such that the acceptance ratio was in the range of $(0.23, 0.29)$. The overall acceptance rate for the MH move in the EE sampler was 0.26. We used 2000 iterations to burn in the EE sampler and then generated 20,000 iterations. Figure 1 shows autocorrelations and the samples generated in each chain based on the last 10,000 iterations. We can see, from Figure 1, that the EE sampler works remarkably well and the high correlations do not impose any difficulty for the EE sampler at all.

8. Example 2. In this example, we consider another extreme and more challenging case, in which we assume a normal mixture distribution with different variances. Specifically, in (7.1) we take

$$\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & \sigma_{i1}\sigma_{i2}\rho_i \\ \sigma_{i1}\sigma_{i2}\rho_i & \sigma_{i2}^2 \end{pmatrix}$$

with $\sigma_{11} = \sigma_{12} = 0.01$, $\sigma_{21} = \sigma_{22} = 1.0$ and $\rho_1 = \rho_2 = 0$. Since the minimum value of the energy function $h(x)$ is around -6.679, we took $H_0 = -7.0$. We first tried the same setting for the energy and temperature ladders with $K = 2$, $p_{\text{ee}} = 0.1$ and the MH proposal $N_2(X_n^{(i)}, \tau_i^2 T_i I_2)$. The chain $X^{(0)}$ was trapped around one mode and did not move from one mode to another at all. A similar result was obtained when we set $K = 4$. So, it did not help to simply increase K. One potential reason for this may be the choice of the MH proposal $N_2(X_n^{(0)}, \tau_0^2 I_2)$ at the lowest energy level. If τ_0 is large, a candidate point around the mode with a smaller variance is likely to be rejected. On the other hand, the chain with a small τ_0 may move more frequently, but the resulting samples will be highly correlated.

Intuitively, an improvement could be made by increasing K, tuning energy and temperature ladders, choosing a better MH proposal and a more appropriate p_{ee}. Several attempts along these lines were made to improve the EE sampler and the results based on one of those trials are given below. In this attempt, K was set to 6,

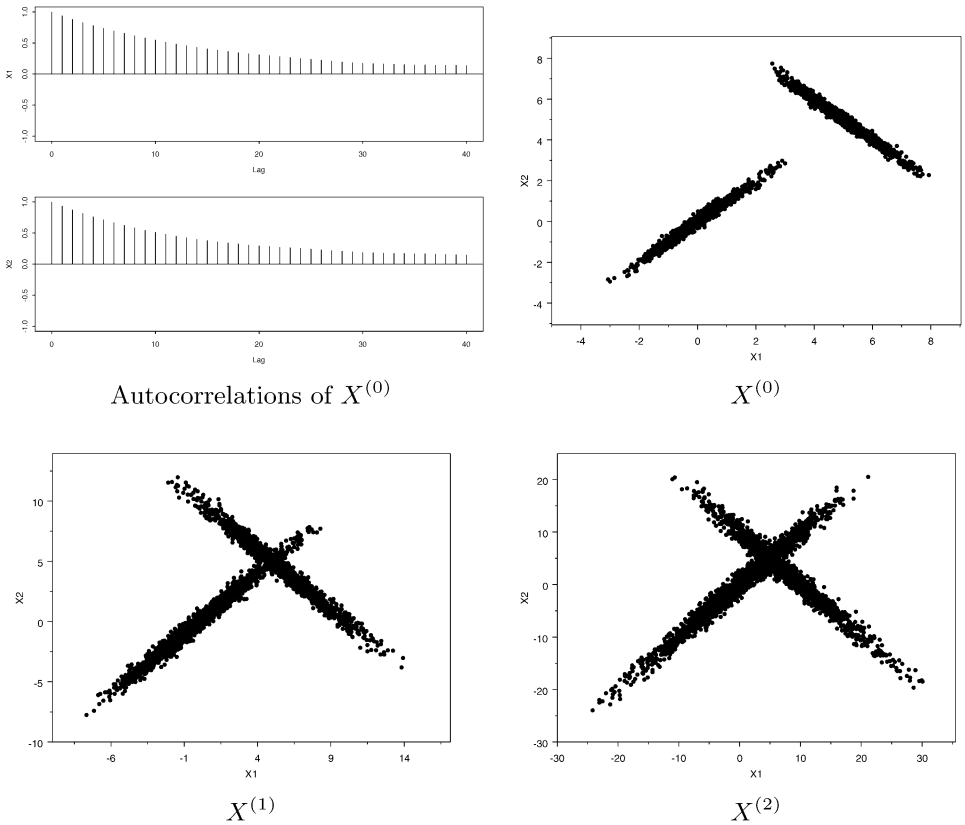

FIG. 1. *Plots of EE samples from a normal mixture distribution with equal variances.*

and $H_1 = \log(4\pi) + \alpha = 2.53 + \alpha$, where α was set to 0.6. The energy ladder was set between H_1 and $H_{\min} + 100$ in a geometric progression, the temperatures were between 1 and 70, and $p_{\text{ee}} = 0.5$. The MH proposals were specified as $N_2(X_n^{(i)}, \tau_i^2 T_i I_2)$ for $i > 0$ and $N_2(\mu(X_n^{(0)}), \Sigma(X_n^{(0)}))$ at the lowest energy level, where $\mu(X_n^{(0)})$ was chosen to be the mode of the target distribution based upon the location of the current point $X_n^{(0)}$ and $\Sigma(X_n^{(0)})$ was specified in a similar fashion as $\mu(X_n^{(0)})$. We used 20,000 iterations to burn in the EE sampler and then generated 50,000 iterations. Figure 2 shows the plots of the samples generated in $X^{(0)}$ based on all 50,000 iterations. The resulting chain had excellent mixing around each mode, and the chain also did move from one mode to another mode. However, the chain did not move as freely as expected.

Due to lack of experience in using the EE sampler, we are not sure at this moment whether the EE sampler can be further improved for this example. If so, we do not know how. We would like the authors to shed light on this.

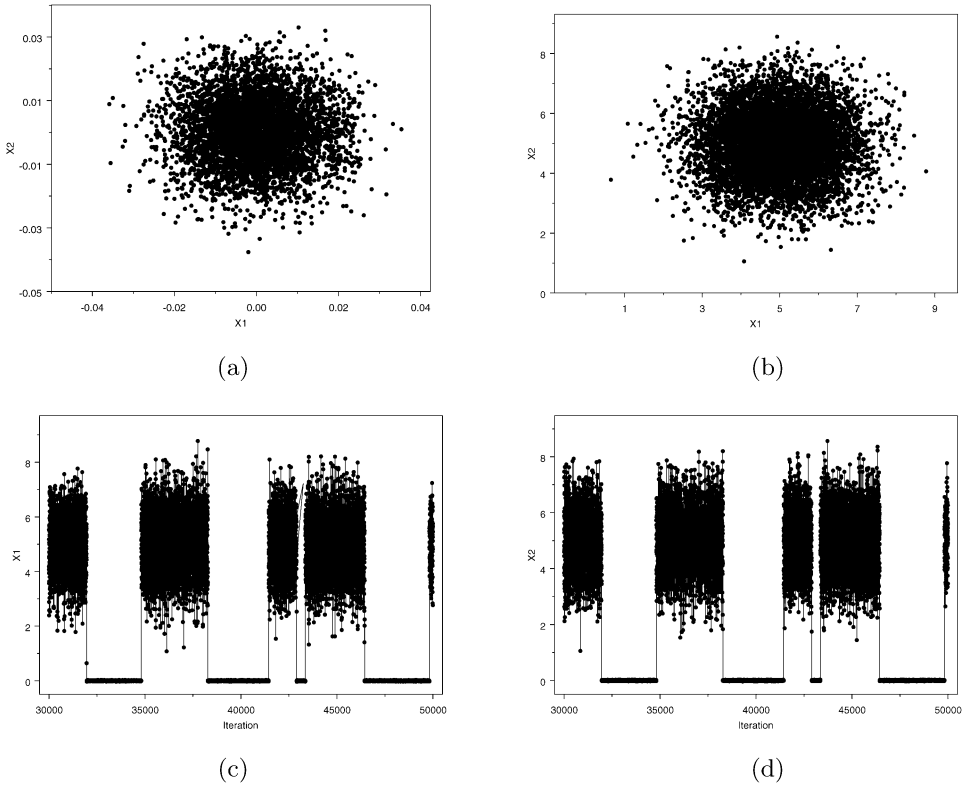

FIG. 2. *Normal mixture distribution with unequal variances. Samples of* $X^{(0)} = (X_1^{(0)}, X_2^{(0)})$ *around* (a) *mode* $(0, 0)$ *and* (b) *mode* $(5, 5)$. *The marginal sample paths of* $X_1^{(0)}$ (c) *and* $X_2^{(0)}$ (d).

9. Discussion. The EE sampler is a potentially useful and effective tool for sampling from a multimodal distribution. However, as shown in Example 2, the EE sampler did experience some difficulty in sampling from a bivariate normal distribution with different variances. For the unequal variance case, the guidelines for practical implementation provided in the paper may not be sufficient. The statement, "the sampler can jump freely between the states with similar energy levels," may not be accurate as well.

As a uniform proposal was suggested for the equi-energy move, it becomes apparent that the points around the modes corresponding to larger variances are more likely to be selected than those corresponding to smaller variances. Initially, we thought that an improvement might be made by assigning a larger probability to the points from the mixand with a smaller variance. However, this would not work as the resulting acceptance probability would become small. Thus, a more likely selected point may be less likely to be accepted. It does appear that a uniform proposal may be a good choice for the equi-energy move.

REFERENCES

[1] CHEN, M.-H., DEY, D. K. and SHAO, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *J. Amer. Statist. Assoc.* **94** 1172–1186. MR1731481

[2] CHEN, M.-H. and SCHMEISER, B. W. (1998). Toward black-box sampling: A random-direction interior-point Markov chain approach. *J. Comput. Graph. Statist.* **7** 1–22. MR1628255

[3] CHEN, M.-H. and SHAO, Q.-M. (2002). Partition-weighted Monte Carlo estimation. *Ann. Inst. Statist. Math.* **54** 338–354. MR1910177

[4] NEAL, R. M. (2003). Slice sampling (with discussion). *Ann. Statist.* **31** 705–767. MR1994729

[5] WU, L. (2004). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *J. Amer. Statist. Assoc.* **99** 700–709. MR2090904

DEPARTMENT OF STATISTICS
UNIVERSITY OF CONNECTICUT
STORRS, CONNECTICUT 06269-4120
USA
E-MAIL: mhchen@stat.uconn.edu
sdkim@stat.uconn.edu

The Annals of Statistics
2006, Vol. 34, No. 4, 1636–1641
DOI: 10.1214/009053606000000470
Main article DOI: 10.1214/009053606000000515

DISCUSSION OF "EQUI-ENERGY SAMPLER" BY KOU, ZHOU AND WONG

BY PETER MINARY AND MICHAEL LEVITT

Stanford University

Novel sampling algorithms can significantly impact open questions in computational biology, most notably the *in silico* protein folding problem. By using computational methods, protein folding aims to find the three-dimensional structure of a protein chain given the sequence of its amino acid building blocks. The complexity of the problem strongly depends on the protein representation and its energy function. The more detailed the model, the more complex its corresponding energy function and the more challenge it sets for sampling algorithms. Kou, Zhou and Wong have introduced a novel sampling method, which could contribute significantly to the field of structural prediction.

1. Rough 1D energy landscape. Most of the energy functions describing off-lattice protein models are assembled from various contributions, some of which take account of the "soft" interactions between atoms (residues) far apart in sequence, while others represent the stiff connections between atoms directly linked together with chemical bonds. As a consequence of this complex nature, the resulting energy function is unusually rough even for short protein chains.

The authors apply the equi-energy (EE) sampler to a multimodal two-dimensional model distribution, which is an excellent test for sampling algorithms. However, it lacks the characteristic features of distributions derived from complex energy functions of off-lattice protein models. In studies conducted by Minary, Martyna and Tuckerman [1], the roughness of such energy surfaces was represented by using a Fourier series on the interval $[0, L = 10]$ [see Figure 1(a)],

$$h(x) = 2\sum_{i=1}^{20} c(i)\sin(i2\pi x/L), \tag{1}$$

where the coefficients are

$$(c_1, c_2, \ldots, c_{20}) = (0.21, 1.25, 0.61, 0.25, 0.13, 0.10, 1.16, 0.18, 0.12, 0.23, 0.21, 0.19, 0.37, 0.99, 0.36, 0.02, 0.06, 0.08, 0.09, 0.04).$$

The performance of various sampling algorithms on the energy function, $h(x)$, is related to their ability to effectively locate the energy basins separated by large

Received November 2005.

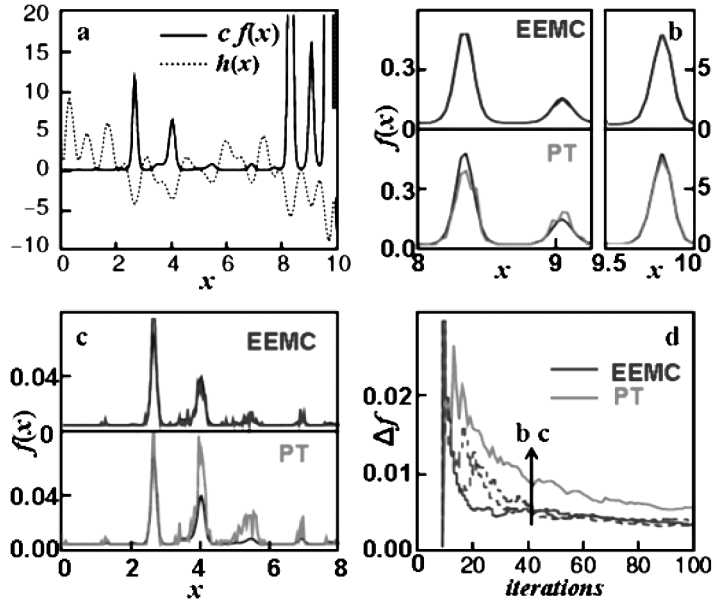

FIG. 1. (a) *The model system energy function*, $h(x)$ (*dotted line*), *and the corresponding normalized distribution*, $f(x)$, *scaled by a constant*, $c = 200$ (*solid line*). (b) *Comparing distributions produced by the EE sampler* (*EEMC*) *and parallel tempering* (*PT*) *to the target distribution* (*black*) *after* 40,000 *iterations in the interval* [0, 8]. (c) *Similar comparison in the intervals* [8, 9.5] *and* [9.5, 10]. (d) *Convergence rate* Δf *to the target distribution* $f(x)$ *as a function of the number of iterations for the EE sampler with energy disk sizes of* 5,000 (*solid black*), 10,000 (*dashed black*) *and* 2,500 (*dot-dashed black*). *The same quantity is plotted for parallel tempering* (*gray*). *The distributions presented in* (b) *and* (c) *are produced from statistics*, *collected up to* 40,000 *iterations* (*arrow*).

energy barriers. In particular, previous studies by Minary, Martyna and Tuckerman [1] show that a superior convergence rate to the corresponding normalized distribution,

$$f(x) = \frac{1}{N}\exp(-h(x)), \qquad N = \int_0^L \exp(-h(x))\,dx, \tag{2}$$

often correlates with enhanced sampling of more complex energy functions.

As a first test, the EE sampler with five Hybrid Monte Carlo chains ($K = 4$) was applied to this problem. Hybrid Monte Carlo (HMC) [2] was used to propagate the chains $X^{(i)}$, as it generates more efficient moves guided by the energy surface gradient. Furthermore, it is well suited to complex high-dimensional systems because it can produce collective moves. The initial values of the chains were obtained from a uniform distribution on $[0, L]$ and the MD step size was finely tuned, so that the HMC acceptance ratio was in the range [0.4, 0.5]. Figure 1 shows that for all $x \in [0, L]$, $h(x) > -10$, so that H_0 was set to -10. The energy levels, which were chosen by geometric progression in the interval $[-10, 10]$, are reported together with the temperature levels in Table 1. The EE jump probability p_{ee} was set

TABLE 1
Sample size of energy rings

	Energy rings				
Chain	**< -8.7**	**$[-8.7, -7.5)$**	**$[-7.5, -5)$**	**$[-5.0, -0.2)$**	**≥ -0.2**
$X^{(0)}$, $T_0 = 1.0$	4295	1981	928	772	24
$X^{(1)}$, $T_1 = 2.0$	2435	1734	1622	3526	683
$X^{(2)}$, $T_2 = 3.9$	726	675	1338	6252	3009
$X^{(3)}$, $T_3 = 7.7$	308	302	895	6847	5648
$X^{(4)}$, $T_4 = 15.3$	240	220	714	7187	7639

to 0.15 and each chain was equilibrated for an initial period prior to the production sampling of 100,000 iterations. The sizes of the energy rings were bounded, as computer memory is limited, especially when applying the EE sampler to structure prediction problems. After their sizes reach the upper bound, the energy rings are refreshed by replacing randomly chosen elements. In Table 1, the number of samples present in each energy ring after the initial burn-in period is summarized. It shows that energy rings corresponding to lower-order chains are rich in low-energy elements, whereas higher-order chains are rich in high-energy elements.

For benchmarking the performance of the EE sampler, parallel tempering (PT) trajectories of the same length were generated using the same number of HMC chains, temperature levels and exchange probabilities. The average acceptance ratio for EE jumps and replica exchange in PT was 0.82 and 0.45, respectively. Figures 1(b) and (c) compare the analytical distributions, $f(x)$, with the numerical ones produced by the EE sampler and PT after 40,000 iterations. All the minima of $f(x)$ are visited by both methods within this fraction of the whole sampling trajectory. Quantitative comparison is obtained via the average distance between the produced and analytical distributions,

$$\Delta f(f_k, f) = \frac{1}{N} \sum_{i=1}^{N} |f_k(x_i) - f(x_i)|, \tag{3}$$

where f_k is the instantaneously computed numerical distribution at the kth iteration and N is the number of bins used. Figure 1(d) depicts Δf, as a function of the number of MC iterations. It is clear that a substantial gain in efficiency is obtained with the EE sampler, although the convergence rate is dependent on the maximum size of energy disks.

2. Off-lattice protein folding in three dimensions. Efficient sampling and optimization over a complex energy function are regarded as the most severe barrier to *ab initio* protein structure prediction. Here, we test the performance of

the EE sampler in locating the native-like conformation of a simplified united-residue off-lattice β-sheet protein introduced by Sorenson and Head-Gordon [4] based on the early works of Honeycutt and Thirumalai [3]. The model consists of 46 pseudoatoms representing residues of three different types: hydrophobic (B), hydrophilic (L) and neutral (N). The potential energy contains bonding, bending, torsional and intermolecular interactions:

$$
\begin{aligned}
h = {} & \sum_{i=2}^{46} \frac{k_{\text{bond}}}{2}(d_i - \sigma)^2 + \sum_{i=3}^{46} \frac{k_{\text{bend}}}{2}(\theta_i - \theta_0)^2 \\
& + \sum_{i=4}^{46} [A(1 + \cos\phi) + B(1 + \cos 3\phi)] \\
& + \sum_{i=1,\, j\geq +3}^{46} V_{XY}(r_{ij}), \qquad X, Y = B, L \text{ or } N.
\end{aligned} \tag{4}
$$

Here, $k_{\text{bond}} = 1000\varepsilon_H$ Å^{-2}, $\sigma = 1$ Å, $k_{\text{bend}} = 20\varepsilon_H\,\text{rad}^{-2}$, $\theta_0 = 105°$; $\varepsilon_H =$ 1000K (Kelvin); the torsional potentials have two types: if the dihedral angles involve two or more neutral residues, $A = 0$, $B = 0.2\varepsilon_H$ (flexible angles), and otherwise $A = B = 1.2\varepsilon_H$ (rigid angles). The nonbonded interactions are bead-pair specific, and are given by $V_{BB} = 4\varepsilon_H[(\sigma/r_{ij})^{12} - (\sigma/r_{ij})^6]$, $V_{LX} = 8/3\varepsilon_H[(\sigma/r_{ij})^{12} + (\sigma/r_{ij})^6]$ for $X = B$ or L and $V_{NX} = 4\varepsilon[(\sigma/r_{ij})^{12}]$ with $X = B$, L or N. This model and its energy function are illustrated in Figure 2.

A particular sequence of "amino acids," $(\text{BL})_2\text{B}_5\text{N}_3(\text{LB})_4\text{N}_3\text{B}_9\text{N}_3(\text{LB})_5\text{L}$, is known to fold into a β-barrel conformation as its global minimum energy structure with the potential energy function given above. Thus, this system is an excellent test of various sampling algorithms such as the EE sampler or parallel tempering. Since the native structure is known to be the global minimum ($h_{\min}$) on the energy surface, H_0 was set to $h_{\min} - 0.05|h_{\min}|$. The energy corresponding to the completely unfolded state (h_{unf}) serves as an approximate upper bound to the energy function because all the favorable nonbonded interactions are eliminated. This is true only if we assume that bond lengths and bend angles are kept close to their ideal values and there are no "high-energy collisions" between nonbonded beads. K was taken to be 8 so that nine HMC chains were employed.

First, the energy levels $H_1, \ldots, H_8$ were chosen to follow a geometric progression in $[H_0, H_{8+1} = h_{\text{unf}}]$, but this produced an average EE jump acceptance ratio of 0.5. In order to increase the acceptance, the condition for geometric progression was relaxed. The following alternative was used: (a) create an energy ladder by using $H_{i+1} = H_i\lambda$; (b) uniformly scale $H_1, \ldots, H_{8+1}$ so that $H_{8+1} = h_{\text{unf}}$. Applying this strategy and using a λ drawn from $[1.1, 1.2]$ produced an average EE jump acceptance ratio of $\sim$0.8. The equi-energy probability p_{ee} was set to 0.15 and the parameters for the HMC chains $X^{(i)}$ were chosen in the same way as discussed in the case of the 1D model problem.

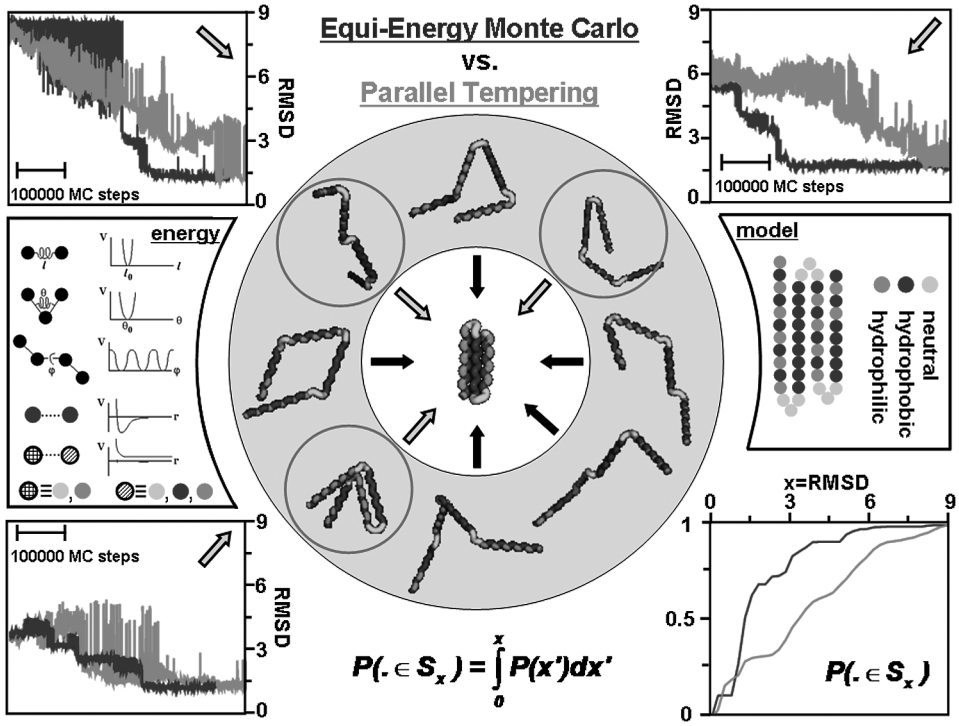

FIG. 2. *Comparing equi-energy Monte Carlo* (*EEMC*) *and parallel tempering* (*PT*) *to fold* 3*D off-lattice* β*-sheet model proteins with known native structure. The figure shows the united-residue model with three types of residues*: *hydrophobic* (*black*), *hydrophilic* (*gray*) *and neutral* (*light gray*). *The energy function contains contributions from bonds*, *bends*, *torsions and intermolecular interactions*, *the last being attractive between hydrophobic–hydrophobic residues and repulsive otherwise. The circular image in the center of the figure illustrates some of the ten initial structures*, *which were generated by randomizing the torsions in the loop regions. These torsions are defined as the ones which include more than two neutral residues. The three "RMSD from native vs. MC steps" subplots contain representative trajectories starting from the three encircled configurations*, *whose distance from the native state* (s_n) *was* $\sim 3.0,\ 6.0$ *and* 9.0 Å, *respectively. The last subplot gives the probability that a visited structure is contained in the set* $S_x = \{s : RMSD(s, s_n) \le x \text{ Å}\}$, *PT* (*gray*) *and EEMC* (*black*).

To test the ability of EEMC and PT to locate the native structure, ten initial structures were obtained by randomly altering the loop region torsion angles. Then both EEMC and PT trajectories starting from the same initial configurations were generated. For each structure (s) the RMSD deviation from the native state (s_n) was monitored as a function of the number of MC iterations. The three representative trajectories depicted in Figure 2 start from initial structures with increasing RMSD distance from the native structure. Some trajectories demonstrate the superior performance of the EE sampler over PT, since the native state is found with fewer MC iterations. More quantitative comparison is provided by the probability

distribution of the RMSD distance, $P(x)$, which was based on a statistic collected from all the ten trajectories. As Figure 2 indicates, the cumulative integral of the distribution shows that 50% of the structures visited by the EE sampler are in $S_{1.5}$ where $S_x = \{s : RMSD(s, s_n) \leq x \text{ Å}\}$. The corresponding number for PT is 25%.

These tests show that the EE sampler can offer sampling efficiency better than that of other state-of-the-art sampling methods such as parallel tempering. Careful considerations must be made when choosing the setting for the energy levels and disk sizes for a given number of chains. Furthermore, we believe that proper utilization of the structural information stored in each energy disk could lead to the development of novel, more powerful topology-based optimization methods.

REFERENCES

[1] MINARY, P., MARTYNA, G. J. and TUCKERMAN, M. E. (2003). Algorithms and novel applications based on the isokinetic ensemble. I. Biophysical and path integral molecular dynamics. *J. Chemical Physics* **118** 2510–2526.

[2] DUANE, S., KENNEDY, A. D., PENDLETON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.

[3] HONEYCUTT, J. D. and THIRUMALAI, D. (1990). Metastability of the folded states of globular proteins. *Proc. Nat. Acad. Sci. USA* **87** 3526–3529.

[4] SORENSON, J. M. and HEAD-GORDON, T. (1999). Redesigning the hydrophobic core of a model β-sheet protein: Destabilizing traps through a threading approach. *Proteins*: *Structure*, *Function and Genetics* **37** 582–591.

DEPARTMENT OF STRUCTURAL BIOLOGY
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: peter.minary@stanford.edu
michael.levitt@stanford.edu

The Annals of Statistics
2006, Vol. 34, No. 4, 1642–1645
DOI: 10.1214/009053606000000506
Main article DOI: 10.1214/009053606000000515

DISCUSSION OF "EQUI-ENERGY SAMPLER" BY KOU, ZHOU AND WONG

BY YING NIAN WU AND SONG-CHUN ZHU

University of California, Los Angeles

We congratulate Kou, Zhou and Wong for making a fundamental contribution to MCMC. Our discussion consists of two parts. First, we ask several questions about the EE-sampler. Then we review a data-driven MCMC scheme for solving computer vision problems.

1. Questions. To simplify the language, we use $\pi(x)$ to denote a distribution we want to sample from, and $q(x)$ to denote a distribution at a higher temperature (with energy truncation). The distributions $\pi(x)$ and $q(x)$ can be understood as two consecutive levels in the EE-sampler. Suppose we have obtained a sample from $q(x)$ by running a Markov chain (with a burn-in period), and let us call the sampled states q-states. Suppose we have also formed the energy rings by grouping these q-states. Now consider sampling from $\pi(x)$ by the EE-sampler.

1. In the jump step, can we make the chain jump to a q-state outside the intended energy ring? For example, can we simply propose to jump to any random q-states, as if performing the Metropolized independent sampler [2] with $q(x)$ as the proposal distribution? Without restricting the jump to the intended energy ring, it is possible that the chain jumps to a q-state of a higher energy level than the current state, but it is also possible that it lands on a lower-energy q-state.

If the energy of the current state is very low, we may not have any q-states in the corresponding energy ring to make the EE jump. But if we do not restrict the chain to the current energy ring, it may jump to a higher-energy q-state and escape the local mode.

The reason we ask this question is that the power of the EE-sampler seems to come from its reuse of the q-states, or the long memory of the chain, instead of the EE feature.

2. Can we go up the distribution ladder in a serial fashion? In the EE-sampler, when sampling $\pi(x)$, the chain that samples $q(x)$ keeps running. Can we run a Markov chain toward $q(x)$ long enough and then completely stop it, before going up to sample $\pi(x)$? What is the practical advantage of implementing the sampler in a parallel fashion, or is it just for proving theoretical convergence?

3. About the proof of Theorem 2, can one justify the algorithm by the Metropolized independent sampler [2], where the proposal distribution is $q(x)$

Received November 2005.

truncated to the current energy range? Of course, there can be a reversibility issue. But in the limit this may not be a problem. In the authors' proof, they also seem to take such a limit. What extra theoretical insights or rigor can be gained from this proof?

4. Can one obtain theoretical results on the rate of convergence? To simplify the situation, consider a Markov chain with a mixture of two moves. One is the regular local MH move. The other is to propose to jump from x to y with $y \sim q$, according to the Metropolized independent sampler. Liu [2] proves that the second largest eigenvalue of the transition kernel of the Metropolized independent sampler is $1 - \min_x \pi(x)/q(x)$. At first sight, this is a discouraging result: even if $q(x)$ captures all the major modes of $\pi(x)$ and takes care of the global structure, the chain can still converge very slowly, because in the surrounding tail areas of the modes, the ratio $\pi(x)/q(x)$ may be very small. In other words, the rate of convergence can be decided by high-energy x that are not important. However, we can regard $q(x)$ as a low-resolution approximation to $\pi(x)$, so we should consider the convergence on a coarsened grid of the state space, where the probability on a coarsened grid point is the sum or integral of probabilities on the underlying finer grid points, so the minimum probability ratio between coarsened π and q may not be very small. The lack of resolution in the above scheme is taken care of by the local MH move. So the two types of moves complement each other to take care of things at two different scales. This seems also the case with the more sophisticated EE sampler.

2. Data-driven MCMC and Swendsen–Wang cut. Similar to EE-sampler, making large jumps to escape local modes is also the motivation for the data-driven (DD) MCMC scheme of Tu, Chen, Yuille and Zhu [3] for solving computer vision problems.

Let $\mathbf{I}$ be the observed image data defined on a lattice Ω, and let W be an interpretation of $\mathbf{I}$ in terms of what is where. One simple example is image segmentation: we want to group pixels into different regions, where the pixel intensities in each region can be described by a coherent generative model. For instance, Figures 1 and 2 show two examples, where the left image is the observed one, and the right

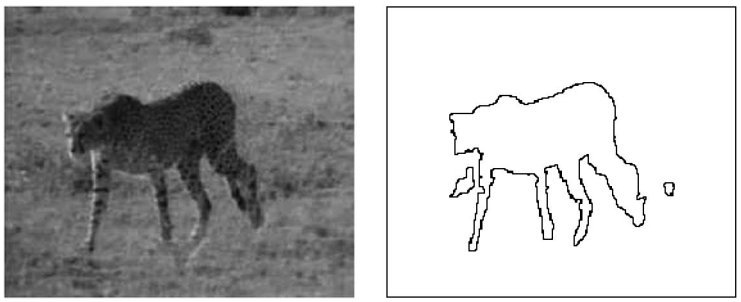

FIG. 1. *Image segmentation.*

FIG. 2. *Image segmentation.*

image displays the boundaries of the segmented regions. Here W consists of the labels of all the pixels: $W = (W_i, i \in \Omega)$, so that $W_i = l$ if pixel i belongs to region $l \in \{1, \ldots, L\}$, where L is the total number of regions.

In a Bayesian formulation, we have a generative model: $W \sim p(W)$ and $[\mathbf{I}|W] \sim p(\mathbf{I}|W)$. Then image interpretation amounts to sampling from the posterior $p(W|\mathbf{I})$. For the image segmentation problem, the prior $p(W)$ can be something like the Potts model, which encourages identical labels for neighboring pixels. The model $p(\mathbf{I}|W)$ can be such that in each region the pixel values follow a two-dimensional low-order polynomial function plus i.i.d. noise.

To sample $p(W|\mathbf{I})$, one may use a random-scan Gibbs sampler to flip the label of one pixel at a time. However, such local moves can be easily trapped in local modes. A DD-MCMC scheme is to cluster pixels based on local image features, and flip all the pixels in one cluster together.

Specifically, for two neighboring pixels i and j, let $p_{i,j} = P(W_i = W_j | F_{i,j}(\mathbf{I}))$, where $F_{i,j}(\mathbf{I})$ is a similarity measure, for example, $F_{i,j}(\mathbf{I}) = |\mathbf{I}_i - \mathbf{I}_j|$. In principle, this conditional probability can be learned from training images with known segmentations. Then for each pair of neighboring pixels (i, j) that belong to the same region under the current state $W = A$, we connect i and j with probability $p_{i,j}$. This gives rise to a number of clusters, where each cluster is a connected graph of pixels. We then randomly pick a cluster V_0 (see Figure 3), and assign a single

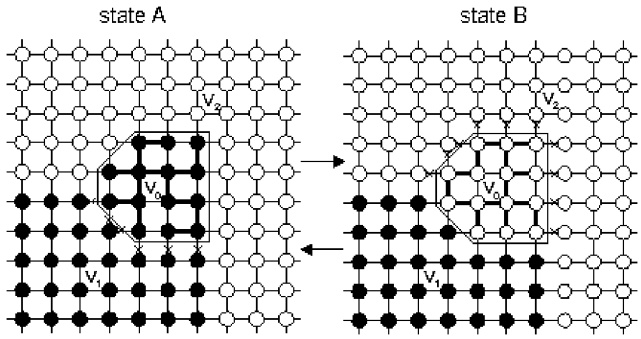

FIG. 3. *Swendsen–Wang cut.*

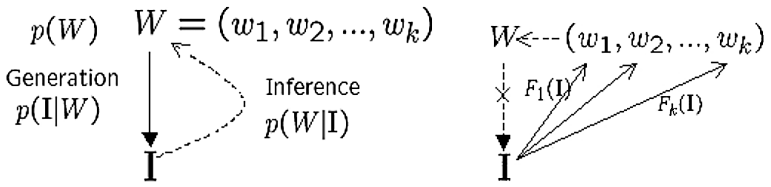

FIG. 4. (a) *Top–down approach*; (b) *bottom-up approach*.

label l to all the pixels in V_0 with probability q_l. One can design q_l so that the move is always accepted, very much like the Gibbs sampler. This is the basic idea of the Swendsen–Wang cut algorithm of Barbu and Zhu [1], which is a special case of DD-MCMC [4]. The algorithm is very efficient. Figures 1 and 2 show two examples where the results are obtained in seconds, thousands of times faster than the single-site Gibbs sampler.

Figure 4 illustrates the general situation for DD-MCMC. Part (a) illustrates the model-based inference, where the top–down generative model $p(W)$ and $p(\mathbf{I}|W)$ is explicitly specified. The posterior $p(W|\mathbf{I})$ is implicit and may require MCMC sampling. Part (b) illustrates the bottom-up operations, where some aspects of W can be explicitly calculated based on some simple image features $\{F_k(\mathbf{I})\}$, without an explicit generative model. The bottom-up approach may not give a consistent and accurate full interpretation W, but it can be employed to design efficient moves for sampling the posterior $p(W|\mathbf{I})$ in the top–down approach. If vision is a bag of bottom-up tricks, then DD-MCMC provides a principled scheme to bag these tricks. The recent work of Tu, Chen, Yuille and Zhu [3] also incorporates boosting into this MCMC scheme.

REFERENCES

[1] BARBU, A. and ZHU, S. C. (2005). Generalizing Swendsen–Wang to sampling arbitrary posterior probabilities. *IEEE Trans. Pattern Analysis and Machine Intelligence* **27** 1239–1253.

[2] LIU, J. S. (1996). Metropolized independence sampling with comparisons to rejection sampling and importance sampling. *Statist. Comput.* **6** 113–119.

[3] TU, Z. W., CHEN, X. R., YUILLE, A. L. and ZHU, S. C. (2005). Image parsing: Unifying segmentation, detection and recognition. *Internat. J. Computer Vision* **63** 113–140.

[4] TU, Z. W. and ZHU, S. C. (2002). Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Trans. Pattern Analysis and Machine Intelligence* **24** 657–673.

DEPARTMENTS OF STATISTICS
AND COMPUTER SCIENCE
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90095
USA
E-MAIL: ywu@stat.ucla.edu
sczhu@stat.ucla.edu

The Annals of Statistics
2006, Vol. 34, No. 4, 1646–1652
DOI: 10.1214/009053606000000524
Main article DOI: 10.1214/009053606000000515

REJOINDER

BY S. C. KOU, QING ZHOU AND WING H. WONG

Harvard University, Harvard University and Stanford University

We thank the discussants for their thoughtful comments and the time they have devoted to this project. As a variety of issues have been raised, we shall present our discussion in several topics, and then address specific questions asked by particular discussants.

1. Sampling algorithms. The widely used state-of-the-art sampling algorithms in scientific computing include temperature-domain methods, such as parallel tempering and simulated tempering, energy-domain methods, such as multicanonical sampling and the EE sampler, and methods involving expanding the sampling/parameter space. The last group includes the Swendsen–Wang type algorithms for lattice models, as Wu and Zhu pointed out, and the group Monte Carlo method [1]. If designed properly, these sampling-space-expansion methods could be very efficient, as Wu and Zhu's example in computer vision illustrated. However, since they tend to be problem-specific, we did not compare the EE sampler with them. The comparison in the paper is mainly between the EE sampler and parallel tempering. Atchadé and Liu's comparison between the EE sampler and the multicanonical sampling thus complements our result. It has been more than 15 years since multicanonical sampling was first introduced. However, we feel that there are still some conceptual questions that remain unanswered. In particular, the key idea of multicanonical sampling is to produce a flat distribution in the energy domain. But we still do not have a simple intuitive explanation of (i) why focusing on the energy works, (ii) why a distribution flat in the energy is sought, and (iii) how such a distribution helps the sampling in the original sample space. The EE sampler, on the other hand, offers clear intuition and a visual picture: the idea is simply to "walk" on the equi-energy sets, and hence focusing on the energy directly helps avoid local trapping. In fact, the numerical results in Atchadé and Liu's comment clearly demonstrate the advantage of EE over multicanonical sampling in the 20 normal mixture example. Specifically, their Table 1 shows that in terms of estimating the probabilities of visiting each mode, the EE sampler is about two to three times more efficient. We think that estimating the probability of visiting individual modes provides a more sensitive measure of the performance, the reason being that even if a sampler misses two or three modes in each run, the sample average of the first and second moments could still be quite

Received December 2005.

good; for example, missing one mode in the far lower left can be offset by missing one mode in the far upper right in the sample average of the first moment, and missing one faraway mode can be offset by disproportionately visiting much more frequently another faraway mode in the sample average of the second moment, and so on. Nevertheless, we agree with Atchadé and Liu that more studies (e.g., on the benchmark phase transition problems in the Ising and Potts models) are needed to reach a firmer conclusion.

2. Implementing the EE sampler for scientific computations. The EE sampler is a flexible and all-purpose algorithm for scientific computing. For a given problem, it could be adapted in several ways.

First, we suggested in the paper that as a good initial start the energy and temperature ladders could be both assigned through a geometric progression. It is conceivable that for a complicated problem alternative assignments might work better, as Minary and Levitt's off-lattice protein folding example illustrated. A good assignment makes the acceptance rates of the EE jump comparably across the different chains, say all greater than 70%. This can be achieved by a small pilot run of the algorithm, which can be incorporated into an automatic self-tuning implementation.

Second, the energy ladder and temperature ladder can be decoupled in the sense that they do not need to always obey $(H_{i+1} - H_i)/T_i \approx c$. For example, for discrete problems such as the lattice phase transition models and the lattice protein folding models, one could take each discrete energy level itself as an energy ring, while keeping the temperatures as a monotone increasing sequence. In this case an EE jump is always accepted, since it always moves between states with the same energy level.

Third, the EE sampler can be implemented in a serial fashion as Wu and Zhu commented. One could start the algorithm from $X^{(K)}$, run for a predetermined number of iterations, completely stop it and move on to $X^{(K-1)}$, run it, completely stop, move on to $X^{(K-2)}$, and so on. This serial implementation offers the advantage of saving computer memory in that one only needs to record the states visited in the chain immediately preceding the current one. The downside is that it will not provide the users the option to online monitor and control (e.g., determine to stop) the algorithm; instead, one has to prespecify a fixed number of iterations to run. In the illustrative multimodal distribution in the paper and the example we include in this rejoinder in Section 4, we indeed utilized the serial implementation since the number of iterations for each chain was prespecified.

Fourth, the EE sampler constructs energy rings to record the footsteps of high-order chains. The fact that a computer's memory is always finite might appear to limit the number of iterations that the EE sampler can be run. But as Minary and Levitt pointed out, this seeming limitation can be readily solved by first putting an upper bound (subject to computer memory) on the energy ring size; once this upper

bound is reached a new sample can be allocated to a specific energy ring by replacing a randomly chosen element in the ring. Minary and Levitt's example involving a rough one-dimensional energy landscape provides a clear demonstration.

Fifth, the key ingredient of the EE sampler is the equi-energy move, a global move that compensates for the local exploration. It is worth emphasizing that the local moves can adopt not only the Metropolis–Hastings type moves, but also Gibbs moves, hybrid Monte Carlo moves as in Minary and Levitt's example, and even moves applied in molecular dynamic simulations, as long as the moves provide good explorations of the local structure.

Sixth, the equi-energy move jumps from one state to another within the same energy ring. As Wu and Zhu commented, it is possible to conduct moves across different energy rings. It has pros and cons, however. It might allow the global jump a larger range, and at the same time it might also lead to a low move acceptance rate, especially if the energy of the current state differs much from that of the proposal jump state. The latter difficulty is controlled in the equi-energy jump of the EE sampler, since it always moves within an energy ring, where the states all have similar energy levels. One way to enhance the global jump range and rein in the move acceptance rate is to put a probability on each energy ring in the jump step. Suppose the current state is in ring D_j. One can put a distribution on the ring index so that the current ring D_j has the highest probability to be chosen, and the neighboring rings D_{j-1} and D_{j+1} have probabilities less than that of D_j to be chosen, and rings D_{j-2} and D_{j+2} have even smaller probabilities to be chosen, and so on. Once a ring is chosen, the target state is proposed uniformly from it.

3. Theoretical issues. We thank Atchadé and Liu for providing a more probabilistic derivation of the convergence of the EE sampler that complements the one we gave in the paper. While these results assure the long-run correctness of the sampler, we agree, however, with Wu and Zhu that investigating the convergence speed is theoretically more challenging and interesting, as it is the rate of convergence that separates different sampling algorithms. So far the empirical evidence supports the EE sampler's promise, but definitive theoretical results must await future studies.

In addition to facilitating the empirically observed fast convergence, another advantage offered by the idea of working on the equi-energy sets is that it allows efficient estimation by utilizing all the samples from *all* the chains on an energy-by-energy basis (as discussed in Section 5 of the paper). We thus believe that the alternative estimation strategy proposed by Chen and Kim is very inefficient, because it essentially wastes all the samples in the chains other than the target one. To make the comparison transparent, suppose we want to estimate the probability of a rare event under the target distribution $P_{\pi_0}(X \in A)$. Chen and Kim's formula would give

$$\hat{P} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{K} w_j \mathbb{1}\big(X_i^{(0)} \in A \cap D_j\big).$$

But since $P_{\pi_0}(X \in A)$ is small, say less than 10^{-10}, there is essentially no sample falling into A in the chain $X^{(0)}$, and correspondingly $\hat{P}$ would be way off no matter how cleverly w_j is constructed. The fact that the high-order chains $X^{(j)}$ could well have samples in the set A (due to the flatness of π_j) does not help at all in Chen and Kim's strategy. But in the EE estimation method such high-order-chain samples are all employed. The tail probability estimation presented in Section 5 and Table 4 illustrates the point. The reason that the EE estimation method is much more efficient in this scenario is due to the well-known fact that in order to accurately estimate a rare event probability importance sampling has to be used and the fact that the EE strategy automatically incorporates importance sampling in its construction. We also want to point out that rare event estimation is an important problem in science and engineering; examples include calculating surface tension in phase transition in physics, evaluating earthquake probability in geology, assessing the chance of bankruptcy in insurance or bond payment default in finance, estimating the potentiality of traffic jams in telecommunication, and so on.

4. Replies to individual discussants. We now focus on some of the individual points raised. Minary and Levitt's discussion has been covered in Sections 1 and 2 of this rejoinder, as was Wu and Zhu's in Sections 1 to 3; we are sorry that space does not permit us to discuss their contributions further.

Atchadé and Liu questioned the derivation of (5) of the paper. This equation, we think, arises directly from the induction assumption, and does not use any assumption on $X^{(i+1)}$ explicitly or implicitly. We appreciate their more probabilistic proof of the convergence theorem.

Chen and Kim asked about the length of the burn-in period in the examples. In these examples the burn-in period consists of 10% to 30% of the samples. We note that this period should be problem-dependent. A rugged high-dimensional energy landscape requires longer burn-in than a smooth low-dimensional one. There is no one-size-fits-all formula.

In the discussion Chen and Kim appeared to suggest that the Gibbs sampler is preferred in high-dimensional problems. But our experience with the Gibbs sampler tells a different story. Though simple to implement, in many cases the Gibbs sampler can be trapped by a local mode or by a strong correlation between the coordinates—the very problems that the modern state-of-the-art algorithms are trying to tackle.

We next consider the needle-in-the-haystack example raised in Chen and Kim's discussion, in which the variances of the normal mixture distribution differ dramatically. Figure 1(a) shows the density function of this example. We implemented the EE sampler using four chains (i.e., $K = 3$) and 200,000 iterations per chain after a burn-in period of 50,000 iterations. Following the energy ladder setting used in Chen and Kim, we set $H_1 = 3.13$; the other energy levels were set between H_1 and $H_{\min} + 100$ $(= 93)$ in a geometric progression: $H_1 = 3.13$, $H_2 = 8.3$, $H_3 = 26.8$.

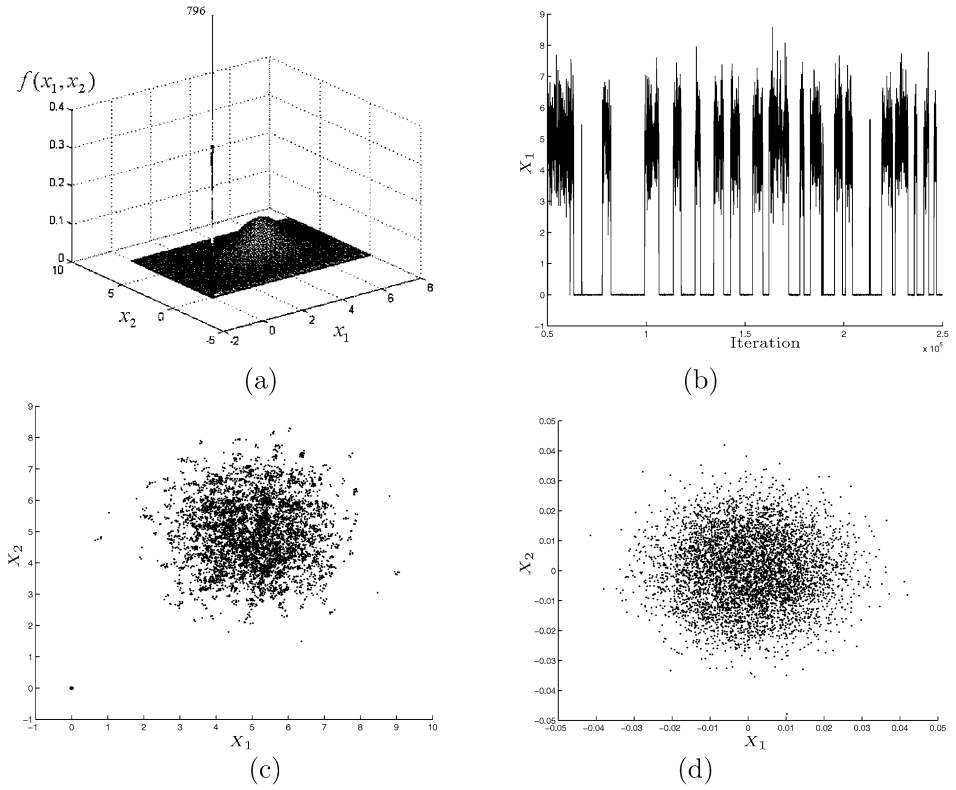

FIG. 1. *The artificial needle-in-the-haystack example.* (a) *The density function of the target distribution.* (b) *The sample path of* $X_1^{(0)}$ *from a typical run of the EE sampler.* (c) *The samples generated at both modes. Note the mode at the origin.* (d) *The samples generated near the mode at the origin.*

The MH proposals were specified as $N_2(\mathbf{X}_n^{(i)}, \tau_i^2 T_i I_2)$, where T_i $(i = 0, \ldots, K)$ is the temperature of the ith chain. We set $\tau_i = 1$ for $i > 0$ and $\tau_0 = 0.05$. The probability of equi-energy jump $p_{\text{ee}} = 0.3$. With all the above parameters fixed in our simulation, we tested the EE sampler with different highest temperatures T_K, whereas the remaining temperatures were evenly distributed on the log-scale between T_K and $T_0 = 1$. We tried T_K =10, 20, 30, 50 and 100; with each parameter setting the EE sampler was performed independently 100 times. From the target chain $\mathbf{X}^{(0)}$ we calculated

$$\hat{P} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(\sqrt{(X_{i1}^{(0)})^2 + (X_{i2}^{(0)})^2} < 0.05\right),$$

the probability of visiting the mode at the origin. From the summary statistics in Table 1, we see that (i) the performance of EE is quite stable with an MSE between 0.04 and 0.06 for different temperature ladders; (ii) more than 98% of the times EE did jump between the two modes. In order to assess the performance of

TABLE 1
Summary statistics of EE and PT for the needle-in-the-haystack example

	$E(\hat{P})$	std($\hat{P}$)	5%	95%	MSE	# Jump	# Miss
EE($N = 200, T_K = 10$)	0.3740	0.2119	0.0289	0.7020	0.0603	36.61	1
EE($N = 200, T_K = 20$)	0.4298	0.2048	0.0556	0.7492	0.0464	40.35	2
EE($N = 200, T_K = 30$)	0.4567	0.1973	0.1188	0.7440	0.0404	43.14	0
EE($N = 200, T_K = 50$)	0.3958	0.2172	0.0223	0.6939	0.0576	39.16	2
EE($N = 200, T_K = 100$)	0.4396	0.2122	0.0986	0.7762	0.0482	39.19	0
EE($N = 50, T_K = 30$)	0.3077	0.3163	0	0.8149	0.1361	6.83	36
PT($N = 200, T_K = 10$)	0.4241	0.2971	0	0.9276	0.0932	364.07	7
PT($N = 200, T_K = 20$)	0.4437	0.2692	0.0000	0.9476	0.0749	157.18	4
PT($N = 200, T_K = 30$)	0.4664	0.3181	0	0.9979	0.1013	104.20	6
PT($N = 200, T_K = 50$)	0.4793	0.3093	0	0.9204	0.0951	63.47	6
PT($N = 200, T_K = 100$)	0.4291	0.2972	0	0.9772	0.0925	36.02	7

Tabulated are the mean, standard deviation, 5% and 95% quantiles, and MSE of $\hat{P}$ in 100 independent runs. Also reported here are the average number of jumps between the two modes and the total number of runs in which the sampler missed the mode at the origin. N is the number of iterations for each chain in units of 1000 after the burn-in period.

EE on this problem, we also applied PT under exactly the same settings including the numbers of chains and iterations, the temperature ladders and the exchange probability ($p_{\text{ex}} = p_{\text{ee}} = 0.3$). It turns out that with all the different temperature ladders PT never outperformed even the worst performance of EE ($T_K = 10$) in MSE (Table 1). From the best performance of the two methods, that is, EE with $T_K = 30$ and PT with $T_K = 20$, one sees that (i) the MSE of EE is about 54% of that of PT; (ii) the spread of the estimated probability is smaller for EE than for PT [see the standard deviation and (5%, 95%) quantiles]. We selected a typical run of EE in the sense that the frequency of jump between the two modes of this run is approximately the same as the average frequency, and we plotted the samples in Figure 1. The chain mixed well in each mode and the cross-mode jump is acceptable. Even in this artificially created extreme example of a needle in the haystack the performance of EE is still quite satisfactory with only four chains ($K = 3$). It is worth emphasizing that we did not even fine-tune the energy or temperature ladders—they are simply set by a geometric progression.

But we do want to point out that one can always cook up extreme examples to defeat any sampling algorithm. For instance, one can hide two needles miles apart in a high-dimensional space, and no sampling algorithm is immune to this type of extreme example. In fact in Chen and Kim's example, if we ran EE with only 50,000 iterations (after the burn-in period) with $T_K = 30$, the resulting MSE increased to 0.136 and 36% of the times EE missed the needle completely (Table 1).

5. Concluding remarks. We thank all the discussants for their insightful contributions. We appreciate the efforts of the Editor and the Associate Editor for putting up such a platform for exchanging ideas. We hope that the readers will enjoy as much as we did reading these comments and thinking about various scientific, statistical and computational issues raised.

REFERENCE

[1] LIU, J. S. and SABATTI, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* **87** 353–369. MR1782484

S. C. KOU
Q. ZHOU
DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
SCIENCE CENTER
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: kou@stat.harvard.edu
zhou@stat.harvard.edu

W. H. WONG
DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: whwong@stanford.edu